UNIVERSIDAD DE CUENCA

Facultad de Ingeniería
Carrera de Ingeniería de Sistemas

Aplicación de técnicas de minería de datos en el contexto del rendimiento académico en la Universidad de Cuenca

Trabajo de titulación previo a la obtención del título de Ingeniero de Sistemas
Modalidad: Proyecto de Investigación

Autor: Cesar Gabriel Loja Rodas
CI.: 0105902456

Director: Ing. Victor Hugo Saquicela Galarza, PhD
CI.: 0103599577

Cuenca-Ecuador

14/10/2019

RESUMEN

Contexto: La Universidad de Cuenca cuenta con una gran cantidad de datos relacionados al rendimiento académico, deserción estudiantil y el estado socio-económico de sus estudiantes que han sido recolectados desde el año 2008 hasta la actualidad, en base a esto es posible extraer información significativa que no está disponible a simple vista, que ayude al cumplimiento de los objetivos estratégicos de desarrollo institucional de la Universidad de Cuenca y de las métricas del modelo de calidad empleado por el Consejo de aseguramiento de la calidad de la educación superior para la evaluación de la educación superior.

Problema: Dentro de este contexto es deseable contar con un análisis del rendimiento y deserción académica, para de esta forma tener un enfoque preventivo que identifique a sujetos en situaciones de riesgo en una etapa temprana de la carrera o ciclo de estudios y que los actores involucrados en estos procesos puedan tomar decisiones mejor informadas basadas en registros históricos.

Solución: El presente trabajo aplica minería de datos para desarrollar modelos de clasificación y predicción, usando la metodología CRISP-DM, se proponen tres problemas específicos a resolver: predicción de deserción, predicción de reprobación por ciclo de de estudios y predicción de reprobación por asignatura.Se pudo concluir que los modelos obtenidos tuvieron un buen porcentaje de aciertos (sobre 73 %) y constituyen una herramienta muy valiosa para la universidad.

Palabras clave: Minería de datos. Rendimiento académico. Deserción

ABSTRACT

Context: The University of Cuenca has a large amount of data related to academic performance, student desertion and the socio-economic status of its students that have been collected since 2008 to the present, based on this is it is possible to extract significant information that is not at first sight, that helps the fulfillment of the strategic objectives of institutional development of the University of Cuenca and of the metrics of the quality model used by the Council for quality assurance of higher education for the evaluation of higher education.

Problem: Within this context it is desirable to have an analysis of academic performance and dropout, in order to have a preventive approach that identifies subjects in situations of risk at an early stage of the career or study cycle and that the actors involved in these processes can make better informed decisions based on historical records.

Solution: This work applies data mining to develop classification and prediction models, using the CRISP-DM methodology, three specific problems are proposed to solve: desertion prediction, prediction of failure by study cycle and prediction of failure by subject. It was possible to conclude that the obtained models had a good percentage of correct answers (about 73 %) and constitute a very valuable tool for the university.

Keywords: Data mining. Academic performance. Dropout

ÍNDICE

RESUMEN 1

ABSTRACT 2

I. INTRODUCCIÓN **12**

1.1. Identificación del problema 12

1.2. Formulación del problema 13

1.3. Justificación 14

1.4. Objetivos y pregunta de la investigación 15

1.4.1. General 15

1.4.2. Específicos 16

1.5. Metodología 17

1.6. Estructura del documento 18

II. MARCO TEÓRICO **19**

2.1. Marco conceptual 19

2.1.1. Minería de Datos 19

2.1.2. Minería de Datos Educativa 21

2.1.3. Metodología Cross Industry Standard Process for Data Mining (CRISP-DM) . . . 22
2.1.4. Conceptos Importantes . . . 24
2.1.4.1. Tomek links . . . 24
2.1.4.2. Normalización . . . 25
2.1.4.3. Valor atípico (outlier) . . . 25
2.1.4.4. Randomized Lasso Method . . . 25
2.1.4.5. K-vecinos más cercanos . . . 26
2.1.4.6. Maquina de soporte de vectores (Support Vector Machine) . . . 26
2.1.4.7. Árbol de decisión (algortimo J48) . . . 26
2.1.4.8. Poda (pruning) . . . 27
2.1.4.9. Bosques aleatorios . . . 27
2.1.4.10. Naive Bayes . . . 27
2.1.4.11. K-means . . . 28
2.1.4.12. Affinity Propagation . . . 28
2.1.4.13. Principal component analysis (PCA) . . . 28
2.2. Revisión Literaria . . . 29
2.2.1. Trabajos relacionados . . . 29
2.2.2. Importancia y necesidad . . . 30
2.2.3. Métodos mas usados . . . 31
2.2.4. Patrones descubiertos . . . 32

III. METODOLOGÍA **33**

3.1. Descripción del dominio de aplicación . . . 33
3.1.1. Objetivos del negocio . . . 33

3.1.2. Datos disponibles 34
3.1.3. Objetivos de la minería de datos 34
3.2. Comprensión de los datos 34
3.2.1. Descripción de los datos 35
3.3. Experimento 1 : Predicción de deserción basado en atributos académicos y socio-económicos relevantes 41
3.3.1. Preparación de los datos 41
3.3.1.1. Datos para el análisis 41
3.3.1.2. Transformación de los datos 42
3.3.1.3. Limpieza y filtrado de datos 44
3.3.1.4. Selección de datos para el análisis 44
3.3.2. Modelado 46
3.3.2.1. Técnicas usadas 47
3.3.2.2. Generación de pruebas de modelo 47
3.3.2.3. Construcción del modelo 48
3.3.2.4. Evaluación del modelo 57
3.4. Experimento 2 : Predicción de reprobación en al menos una asignatura, basado en rendimiento académico pasado. 59
3.4.1. Preparación de los datos 59
3.4.1.1. Datos para el análisis 59
3.4.1.2. Transformación de los datos 59
3.4.2. Limpieza y filtrado de los datos 60
3.4.2.1. Selección de datos para el análisis 61
3.4.3. Modelado 62

3.4.3.1. Técnicas usadas 62
3.4.3.2. Generación de pruebas modelo 63
3.4.3.3. Construcción del modelo 63
3.4.3.4. Evaluación del modelo 66
3.5. Experimento 3 : Predicción de reprobación de asignaturas especificas, al momento de matriculación, basado en el datos académicos históricos 69
3.5.1. Preparación de los datos 69
3.5.1.1. Datos para el análisis 69
3.5.1.2. Transformación de los datos 69
3.5.1.3. Limpieza y filtrado de los datos 72
3.5.1.4. Selección de datos para el análisis 73
3.5.2. Modelado 74
3.5.2.1. Técnicas a usar 74
3.5.2.2. Generación de pruebas modelo 74
3.5.2.3. Construcción del modelo 74
3.5.2.4. Evaluación del modelo 76
3.6. Despliegue 77
3.6.1. Arquitectura del sistema de recomendación 77

IV. DISCUSIÓN DE RESULTADOS **80**

V. CONCLUSIONES **82**

5.1. Conclusiones 82
5.2. Futuras líneas de investigación 84

REFERENCIAS **88**

Índice de tablas

1. Número de registros almacenados en base de datos 35
2. Tablas relacionadas a la matrícula del estudiante 36
3. Tablas relacionadas a la aprobación de asignaturas 37
4. Tablas relacionadas a las asignaturas de la universidad 38
5. Tablas relacionadas a oferta de asignaturas 39
6. Campos de la ficha socio-económica 40
7. Datos de entrada para el modelado 43
8. Atributos seleccionados para el modelado 46
9. Resultado del algoritmo Perceptron multicapa con parámetros óptimos 50
10. Resultado del algoritmo k-vecinos mas cercanos con parámetros óptimos . . . 54
11. Resultado del algoritmo bosques aleatorios sin sobre-ajuste 56
12. Atributos seleccionados para el modelado 62
13. Atributos obtenidos mediante PCA 62
14. Puntaje de los tres métodos de clusterizacion 67
15. Conjunto de datos luego de transformación e integración 72
16. Conjunto de datos luego de limpieza de datos 73

Índice de figuras

1. Diferencia de número de matrículas entre periodos para desertores. 40
2. Ranking de promedio de variables 45
3. Exactitud de los 8 algoritmos iniciales 48
4. Influencia del número de capas en la exactitud del modelo 49
5. Matriz de confusión de algoritmo Perceptron multicapa 50
6. Curva de aprendizaje algoritmo Perceptron multicapa 51
7. Influencia del número de vecinos en la exactitud del modelo 52
8. Matriz de confusión del algoritmo k-vecinos mas cercanos 53
9. Curva de aprendizaje del algoritmo k-vecinos mas cercanos 53
10. Influencia de la profundidad máxima en la exactitud del modelo 55
11. Curva de aprendizaje del algoritmo bosques aleatorios 56
12. Distribución de puntajes de los 3 algoritmos 57
13. Distribución de puntajes de los 3 algoritmos con variables académicas 58
14. Outliers eliminados 60
15. Elección del número de clusters para k-means 64
16. Visualización de datos agrupados con k-meas y 8 clusters. 64
17. Visualización de datos agrupados con AffinityPropagation. 65

19. Porcentaje de reprobados k-means 66
18. Porcentaje de reprobados Propagation affinity 31 clusters 66
21. Prueba t pareada para el número minino de muestras por hoja 75
22. Prueba t pareada para el factor de confianza 76
23. Comparación entre J48 y Naive Bayes 76
24. Reporte de clasificación algoritmo j48 77
25. Flujo de datos para presentación de resultados 78
26. Interfaz aplicación prototipo 79

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

Cesar Gabriel Loja Rodas en calidad de autor y titular de los derechos morales y patrimoniales del trabajo de titulación "Aplicación de técnicas de minería de datos en el contexto del rendimiento académico en la Universidad de Cuenca", de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca,14 de octubre de 2019

Cesar Gabriel Loja Rodas

C.I: 0105902456

Cláusula de Propiedad Intelectual

Cesar Gabriel Loja Rodas, autor del trabajo de titulación "Aplicación de técnicas de minería de datos en el contexto del rendimiento académico en la Universidad de Cuenca", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autor/a.

Cuenca, 14 de octubre de 2019

Cesar Gabriel Loja Rodas

C.I: 0105902456

Capítulo I

INTRODUCCIÓN

1.1. Identificación del problema

El Consejo de aseguramiento de la calidad de la educación superior (CACES) ejecuta procesos de evaluación quinquenal con fines de acreditación a todas las universidades y escuelas politécnicas del Sistema de Educación Superior ecuatoriano, para determinar la categorización de las instituciones según lo determina la Ley Orgánica de Educación Superior – LOES –(CACES, 2018a). Dentro del modelo de evaluación institucional, se destaca un criterio que se refiere a los estudiantes, en donde se consideran las políticas y acciones emprendidas por la institución para garantizar y promover condiciones adecuadas que permitan a los estudiantes alcanzar resultados exitosos en su carrera académica (CACES, 2018b). Dentro de este proceso para evaluar la eficacia de las acciones emprendidas por la universidad en referencia al criterio antes mencionado, se usa un indicador que mide la tasa de retención estudiantil. Diferentes estudios buscan establecer las causas principales de una baja retención, por ejemplo, Ruiz y cols. (2009) indican que los principales factores que influyen son: rendimiento académico, capacidad económica y

orientación vocacional. Específicamente, el rendimiento académico es un indicador de eficacia y calidad educativa, las manifestaciones de fracaso como el bajo rendimiento académico, la repetición y la deserción, expresan deficiencias en un sistema universitario. Las causas del fracaso estudiantil se deben buscar más allá del estudiante mismo, puesto que éste no es el único responsable de su fracaso, lo es también la institución educativa.

1.2. Formulación del problema

Dentro de este contexto es importante que la institución realice un diagnostico permanente del rendimiento académico, pues esto permite, por una parte, la prevención del fracaso y por otra, el "tratamiento" para combatir el fracaso. Plantear un enfoque preventivo, significa tomar en consideración que la intervención educativa debe llevarse a cabo antes de que se haya alcanzado la situación de fracaso escolar. Esto supone que un enfoque preventivo desde la universidad debe, por un lado, identificar a los estudiantes en riesgo de fracaso escolar, y por otro, identificar las situaciones que conducen a que se presente o incremente este riesgo (Artunduaga Murillo, 2008). El problema radica en que las autoridades universitarias no cuentan actualmente con la información que permita realizar un diagnóstico del rendimiento de los estudiantes y así poder tomar decisiones basadas en datos reales.

Es por esto, que en este trabajo se propone la realización de un análisis del rendimiento académico a través de técnicas de Minería de Datos tomando en cuenta los diferentes parámetros que indica Artunduaga Murillo (2008), siempre y cuando estos parámetros estén disponibles y sean relevantes para el análisis y construcción del modelo. Entre los criterios que se indican están incluidos los tipo contextual, tales como, socioculturales, institucionales y pedagógicos, y otros de tipo personal que incluyen demográficos, cognoscitivos y actitudinales, considerando

que existen ciertas variables de tipo estructural que son difíciles de modificar a través de la intervención educativa, como las variables socioculturales y demográficas.

El análisis se realizará mediante el desarrollo de uno o varios modelos de ayuda a la toma decisiones, tanto para los estudiantes, como para las autoridades de las instituciones educativas, enfocados a la predicción del rendimiento y deserción académica; basado en diferentes técnicas de minería de datos, como, clasificación y regresión (Riquelme Santos y cols., 2006).

Con la finalidad de que este proceso se lleve de una forma ordenada, se va a emplear la metodología CRISP-DM [1], que enmarca todas las fases para el desarrollo de un proyecto de minería de datos: entendimiento del dominio y los datos, preparación de los datos, modelado, evaluación y despliegue (Vialardi y cols., 2011), luego de completar las tareas del proceso de descubrimiento de conocimiento descritas, se obtendrá un conjunto de reglas o patrones que formaran un modelo del rendimiento académico de los estudiantes, clasificándolos en diferentes categorías. De esta manera se pretende proveer de una herramienta que permita el soporte de la decisiones que tomen autoridades y estudiantes en lo que respecta al desarrollo de su carrera .

1.3. Justificación

Además de estar presentes dentro del modelo de calidad empleado por el CACES, el rendimiento académico y deserción estudiantil se encuentran ligados a los objetivos estratégicos de desarrollo institucional de la Universidad de Cuenca, en los cuales se destaca, incrementar la tasa de titulación de grado y posgrado(Universidad de Cuenca, 2017), de la misma forma se tiene como indicador a la tasa de retención de grado, la cual hace alusión al nivel de perma-

[1] CRoss Industry Standard Process for Data Mining, en 1997 se puso en marcha como un proyecto de la Unión Europea bajo la iniciativa de financiación ESPRIT. El proyecto fue dirigido por cinco empresas: SPSS, Teradata, Daimler AG, NCR y Ohra

nencia e indirectamente el nivel de deserción de los estudiantes de la institución al inicio de su carrera. Ahí se menciona también que la institución implementa procesos académicos que garantizan la permanencia de los estudiantes en sus estudios. El porcentaje de retención esperado es de 80 % (Universidad de Cuenca, 2017). En este contexto, en este trabajo se pretende aportar con mecanismos mediante los cuales estos objetivos e indicadores se puedan cumplir de una mejor manera. Estos mecanismos incluyen modelos de recomendación y ayuda en la toma decisiones, enfocados tanto a las autoridades académicas como a los estudiantes al momento de su matriculación.

La Universidad de Cuenca cuenta con una gran cantidad de datos que han sido recolectados a través de los años por sus diferentes sistemas de información. La parte central de este trabajo es extraer conocimiento de estos registros mediante la creación de modelos de clasificación y predicción que guíen a los estudiantes y autoridades en los diferentes procesos académicos de la institución como la matriculación y oferta de asignaturas; de forma que los actores involucrados en estos procesos puedan tomar decisiones mejor informadas basadas en registros históricos. Los sistemas que usan estos modelos se basan en la idea de que las personas con el mismo perfil generalmente tienen preferencias similares y, a menudo, toman las mismas decisiones. En la mayoría de los casos, son bien aceptados por los usuarios y ofrecen buenos resultados en una gran variedad de aplicaciones(Vialardi y cols., 2011).

1.4. Objetivos y pregunta de la investigación

1.4.1. General

Aplicar técnicas de Minería de Datos en el contexto del rendimiento académico y deserción estudiantil, con el fin de proveer a las autoridades académicas y estudiantes de un modelo de

ayuda para la toma de decisiones y con esto poder tomar acciones dirigidas a reducir la tasa de deserción estudiantil.

1.4.2. Específicos

- Realizar un proceso de análisis de variables y fuentes de datos que intervienen en el rendimiento académico.
- Obtener varios modelos, conformados por reglas o patrones que permitan la clasificación o predicción de estudiantes en diferentes categorías de rendimiento académico o estado académico (cursando o deserto).
- Validar la precisión de los modelos generados. item Proponer un sistema prototipo que permita la visualización de datos y predicciones.

En razón a los objetivos propuestos anteriormente, se definieron las siguientes preguntas de investigación que enmarcan el presente trabajo:

Pregunta 1: ¿Se puede predecir si un estudiante va a egresar o abandonar sus estudios, con una precisión razonable, en una etapa temprana de su carrera; basado en sus calificación y/o situación socio-económica?

Pregunta 2: ¿Se puede predecir si un estudiante va a reprobar al menos una asignatura, con una precisión razonable, al inicio de cada periodo académico; basado en sus calificación anteriores?

Pregunta 2: ¿Se puede predecir si un estudiante va a reprobar una asignatura especifica, con una precisión razonable, al inicio de cada periodo académico; basado en sus calificación anteriores?

1.5. Metodología

Para el proceso de minería de datos, se investigó sobre las metodologías mas usadas en el campo educativo, se encontró que la mas usada en este tipo de trabajos es la Metodología Cross Industry Standard Process for Data Mining (CRISP-DM), que cuenta con las etapas de: entendimiento del negocio, donde se establecieron los objetivos de la universidad de Cuenca en relación al rendimiento académico y se establecieron los objetivos de la minería de datos; entendimiento de los datos, en el cual se identificaron las fuentes de datos relevantes para el trabajo y su significado; preparación y transformación de datos, se aplicaron procesos para la selección de atributos relevantes, limpieza de datos faltantes, eliminación de outliers, formateo de datos para las herramientas de modelado y cálculo de atributos personalizados; modelado, aquí se aplicaron algoritmos de clasificación, clusterizacion y arboles de decisión; evaluación, se validaron los resultados obtenidos, contrastándolos con los objetivos establecidos en la primera parte; y despliegue, en el que se especifica la forma en la que serán presentados los modelos a los actores interesados.

Se aplicó esta metodología para responder las preguntas de investigación mediante la elaboración de tres experimentos: predicción de deserción estudiantil basado en rendimiento académico y factores socio-económicos, predicción de reprobación de asignaturas en un periodo de estudios y predicción de probabilidad de reprobación en asignaturas especificas al momento de la matriculación.

Para la construcción de los diferentes modelos de predicción se usaron las herramientas de Python y librerías orientadas al manejo y minería de datos, como son: pandas, sklearn y

matplotlib, así también la herramienta WEKA que permite con facilidad la clasificación de registros en base a atributos categóricos, además para el pre-procesamiento de datos se empleo la herramienta de Pentaho, en la que se construyeron procesos de extracción, transformación y carga.

1.6. Estructura del documento

En esta sección se presenta una breve descripción de los capítulos desarrollados en esta tesis.

El **capítulo 2** , Marco Teórico, presenta los conceptos generales sobre minería de datos, su uso en el ámbito educativo y las diferentes técnicas usadas, se presenta de la misma manera una descripción detallada de la metodología de minería de datos CRISP-DM, aquí también se analizan diferentes investigaciones realizadas en el ámbito, se describe su importancia, cuales fueron los métodos mas usados y que patrones se obtuvieron.

El **capítulo 3**, Metodología, se aplica el método CRISP-DM para el caso de estudio de la Universidad de Cuenca, se describe los objetivos del negocio y de la minería de datos, se analizan las fuentes de datos disponibles y se eligen los atributos mas relevantes; se aplican los algoritmos de predicción y se valida sus resultados, finalmente se propone una forma de despliegue y presentación de resultados.

El **capítulo 4**, Discusión de Resultados, se analiza las métricas obtenidas por los experimentos anteriores y los procesos ejecutados para su optimización

El **capítulo 5**, Conclusiones, se presenta una discusión a partir de los resultados obtenidos, donde se considera si se cumplieron los objetivos y si se respondieron a las preguntas de investigación plantadas, se evalúa también el aporte de los modelos generados, en el contexto de las necesidades de la institución. Además, se determinan futuras lineas de investigación,

Capítulo II

MARCO TEÓRICO

2.1. Marco conceptual

2.1.1. Minería de Datos

Debido a la gran cantidad de datos recolectados a lo largo de los años en sistemas de información y los que se siguen generando día a día, los humanos carecemos de la capacidad extraer conocimiento sin depender de herramientas computacionales. Como resultado de esto, como indica Han (2011), en muchas ocasiones estos datos se convierten en una suerte de "tumbas", puesto que los datos están presentes, pero son rara vez visitados al momento de tomar decisiones de negocio. Por lo tanto, las decisiones importantes no se toman en base a la historia que cuentan los repositorios de datos, sino más bien en la intuición de quien toma las decisiones, simplemente porque el que toma las decisiones no tiene las herramientas para extraer el conocimiento incorporado en el gran volumen de datos existente. Es por esto que Han (2011) llama al desarrollo de herramientas que permitan de minería de datos que conviertan estas "tumbas.[en] "minas de oro"del conocimiento.

La minería de datos se define como el proceso de descubrir patrones en los datos. El proceso debe ser automático o (más habitualmente) semiautomático. Los patrones descubiertos deben ser significativos, ya que conducen a alguna ventaja, generalmente una ventaja económica. Los datos están invariablemente presentes en cantidades sustanciales. Los patrones útiles permiten hacer predicciones no triviales sobre nuevos datos (Witten, 2005).

Si bien los términos Knowledge Database Discovery, o Descubrimiento de Conocimiento en Bases de Datos (KDD) y minería de datos están entrelazados como se puede notar en el concepto de minería de datos descrito anteriormente, el concepto de descubrimiento ocupa un papel central, están tan entrelazados que muchos autores tratan a la minería de datos como un sinónimo del término Descubrimiento de Conocimiento en Bases de Datos o Knowledge Database Discovery (KDD), aunque hay otros que según Han (2011), toman a la minería de datos como un subproceso de KDD. Debido a su cercanía en términos conceptuales, se puede tomar como referencia el proceso de KDD para la realización de procesos de minería de datos. El proceso de KDD como lo describe Han (2011) consta de 7 pasos:

1. Limpieza de datos (remover ruido y datos inconsistentes)
2. Integración de datos (combinar múltiples fuentes de datos)
3. Selección de datos (extraer datos relevantes para el análisis)
4. Transformación de datos (formatear los datos para los algoritmos de minería)
5. Minería de datos (proceso para extraer patrones)
6. Evaluación de patrones (identificar los patrones que mejor describen los datos)
7. Presentación del conocimiento (representar patrones y datos de forma visual)

2.1.2. Minería de Datos Educativa

La minería de datos educativa (EDM) es un campo que explota algoritmos de: estadística, aprendizaje automático y minería de datos, sobre diferentes tipos de datos educativos . Su principal objetivo es analizar estos tipos de datos con el fin de resolver temas de investigación y práctica educativa. El proceso de EDM convierte los datos sin procesar de los sistemas educativos en información útil que podría tener un gran impacto en la institución. Este proceso no difiere mucho de otras áreas de aplicación de minería de datos, como negocios, genética, medicina, etc., porque sigue los mismos pasos que el proceso general de minería de datos (Romero y Ventura, 2010).

En Romero y Ventura (2010) se presenta un categorización basada en los objetivos de 300 diferentes investigaciones relacionadas con EDM, las cuatro tareas principales incluyen: análisis y visualización de datos cuyo objetivo es resaltar información útil para la toma de decisiones, la estadística y visualización de información son las dos técnicas mas usadas, la siguiente tarea es proveer retroalimentación para dar soporte a instructores, como fin tiene proveer conocimiento sobre como mejorar el aprendizaje de los estudiantes y como administrar de mejor manera los recursos educativos, esto permite tomar acciones pro activas y remediales a favor de los estudiantes, como técnicas principales se tiene a: las reglas de asociación, clusters, clasificación, análisis de patrones secuenciales, modelado de dependencias y predicción , a continuación se tiene a los sistemas de recomendación para estudiantes, su finalidad es dar información útil directamente a los estudiantes sobre sus actividades personales, como la siguiente asignatura a tomar, el número de créditos por periodo o la secuencia en la que debe seguir sus materias; varias técnicas han sido usadas para esta tarea, las mas comunes son, minería de reglas de asociación, clusters y minería de patrones secuenciales, finalmente se tiene la tarea de predicción del rendimiento de estudiantes, su objetivo es estimar un valor desconocido de una variable

que describe al estudiante, generalmente en el ámbito educativo estas variables describen el rendimiento, conocimiento o notas. Estos valores pueden ser tanto numéricos/continuos como categóricos/discretos. Predicción del rendimiento del estudiante es una de las aplicaciones mas antiguas y populares de EDM, y diferentes técnicas y modelos han sido aplicados: redes neuronales, redes bayesianas, sistemas basados en reglas, regresión y análisis de correlaciones.

2.1.3. Metodología Cross Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM es una de las metodologías para procesos de minería de datos más usadas para la implementación de proyectos de este tipo, dentro de la industria es considerada la metodología de facto. Esta metodología describe las actividades que se deben realizar para desarrollar un proyecto de minería de datos. Cada actividad se compone de tareas, para cada tarea se detallan las salidas generadas y las entradas necesarias. CRISP-DM surge para resolver los problemas que existían en los desarrollos de proyectos de minería de datos. Sus principales objetivos se detallan a continuación(Mariscal y cols., 2010):

- Asegura la calidad de los resultados
- Reduce las habilidades requeridas para la minería de datos
- Captura experiencias para su reutilización
- Propósito general
- Robusto
- Independiente de técnicas y herramientas

- Adaptable a herramientas

El modelo del proceso de CRISP-DM consta de seis fases. Su secuencia no es estricta, es posible moverse entre las diferentes fases. Depende del resultado de la fase anterior, la tarea que se debe realizar a continuación. A continuación se presenta un resumen de cada una de las fases (Mariscal y cols., 2010) :

- Descripción del dominio de aplicación, la fase inicial se enfoca en el entendimiento de los objetivos del proyecto y requerimientos desde una perspectiva del negocio, y para luego convertir este conocimiento en la definición del problema de minería de datos, y el plan preliminar diseñado para alcanzar los objetivos.

- Comprensión de datos, la fase de compresión de datos comienza con una recopilación de datos inicial y continúa con las actividades para familiarizarse con los datos, identificar los problemas de calidad de los datos, descubrir las primeras percepciones de los datos o detectar subconjuntos interesantes para formar hipótesis en base a información oculta.

- Preparación de datos, esta fase cubre todas las actividades relacionadas con la construcción del dataset final , partiendo desde los datos iniciales sin procesar. Es probable que las tareas de preparación de datos se realicen varias veces, y no en un orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de datos para herramientas de modelado.

- Modelado, en esta fase, varias técnicas de modelos son seleccionadas y aplicadas, y sus parámetros son calibrados para obtener valores óptimos. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos sobre la forma de los datos. Por lo tanto, a menudo es necesario volver a la fase de preparación de datos.

- Evaluación, antes de proceder con el despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para la construcción del modelo, además hay que asegurarse que cumpla con los objetivos de negocio. Un objetivo clave de esta fase es determinar si existe algún problema de negocio importante que no se haya considerado suficientemente. Al final de esta fase, se debe tomar una decisión sobre el uso de los resultados de la extracción de datos.

- Despliegue, generalmente la creación del modelo no es el final del proyecto. Incluso si su propósito es aumentar el conocimiento de los datos, será necesario organizar los conocimientos extraídos, así como presentarlos de manera útil al cliente. Dependiendo de los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible. En muchos casos, será el cliente, no el analista de datos, quien llevará a cabo los pasos de implementación. Sin embargo, incluso si el analista no lleva a cabo el esfuerzo de implementación, es importante que el cliente entienda de antemano qué acciones deberán llevarse a cabo para poder utilizar los modelos creados.

2.1.4. Conceptos Importantes

2.1.4.1. Tomek links

Los enlaces de Tomek eliminan la superposición no deseada entre las clases en las que se eliminan los enlaces de la clase mayoritaria hasta que todos los pares vecinos más próximos con una distancia mínima sean de la misma clase (Zhu y cols., 2014). Este método es usado para balanceo de clases.

2.1.4.2. Normalización

Normalización o escalado de atributos es una fórmula, que es usada para escalar los datos, cuando atributos diferentes tienen una alta diferencia es sus valores máximos y mínimos.Esto es un requisito común para muchos estimadores de aprendizaje automático: pueden comportarse mal si las características individuales no se parecen más o menos a los datos estándar normalmente distribuidos (Sieminski y cols., 2018).

2.1.4.3. Valor atípico (outlier)

Un valor atípico es una observación sorprendentemente alejada de algún valor central. Es un valor inusual en relación con la mayor parte de los datos. Las cantidades comúnmente calculadas, como los promedios y las líneas de mínimos cuadrados, pueden verse afectadas drásticamente por tales valores. Se necesitan métodos para detectar valores atípicos y para moderar sus efectos (Tukey, 1977).

2.1.4.4. Randomized Lasso Method

Lasso es es un método de análisis de regresión que realiza selección de variables y regularización para mejorar la exactitud e interpretabilidad del modelo estadístico producido por este. Randomized Lasso funciona mediante el nuevo muestreo de los datos de entrenamiento y el cálculo de un LASSO en cada nuevo muestreo. En resumen, las características seleccionadas más a menudo son buenas características. También se conoce como selección de estabilidad (Meinshausen y Bühlmann, 2010).

2.1.4.5. K-vecinos más cercanos

El modelo de predicción k vecino más cercano (knn) almacena todo el conjunto de datos, y como su nombre lo indica, para predecir una nueva observación, el predictor encuentra las k observaciones en los datos de entrenamiento con vectores de características cercanos a los que deseamos predecir el resultado. La predicción depende de la función de pérdida (Ye, 2003).

2.1.4.6. Maquina de soporte de vectores (Support Vector Machine)

Un SVM es un algoritmo que funciona de la siguiente manera. Utiliza un mapeo no lineal para transformar los datos de entrenamiento originales en una dimensión más alta. Dentro de esta nueva dimensión, busca el hiperplano de separación óptimo lineal (es decir, un "límite de decisión"que separa las tuplas de una clase de otra). Con una asignación no lineal apropiada a una dimensión suficientemente alta, los datos de dos clases siempre pueden estar separados por un hiperplano. La SVM encuentra este hiperplano utilizando vectores de soporte (tuplas de entrenamiento esenciales) y márgenes (definidos por los vectores de soporte) (Han, 2011).

2.1.4.7. Árbol de decisión (algortimo J48)

Un árbol de decisión es una estructura de árbol similar a un diagrama de flujo, donde cada nodo interno (nodo no hoja) denota una prueba en un atributo, cada rama representa un resultado de la prueba y cada nodo hoja (o nodo terminal) tiene una etiqueta de clase . El nodo superior de un árbol es el nodo raíz.

Dada una tupla, X, para la cual se desconoce la etiqueta de clase asociada, los valores de los atributos de la tupla se comparan con el árbol de decisión. Se rastrea una ruta desde la raíz a un nodo hoja que contiene la predicción de clase para esa tupla. Los árboles de decisión se pueden convertir fácilmente en reglas de clasificación (Han, 2011).

2.1.4.8. Poda (pruning)

Cuando se construye un árbol de decisión, muchas de las ramas reflejarán anomalías en los datos de entrenamiento debido al ruido o los valores atípicos. Los métodos de poda de árboles resuelven este problema de sobre-ajuestar los datos. Estos métodos suelen utilizar medidas estadísticas para eliminar las ramas menos confiables (Han, 2011).

2.1.4.9. Bosques aleatorios

Cada uno de los clasificadores en el conjunto es un de árbol de decisión, de modo que la colección de clasificadores es un "bosque". Los árboles de decisión individuales se generan utilizando una selección aleatoria de atributos en cada nodo para determinar la división. De manera más formal, cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles en el bosque. Durante la clasificación, cada árbol vota y se devuelve la clase más popular (Han, 2011).

2.1.4.10. Naive Bayes

Los clasificadores bayesianos son clasificadores estadísticos. Pueden predecir las probabilidades de pertenencia a una clase, como la probabilidad de que una tupla determinada pertenezca a una clase en particular. La clasificación bayesiana se basa en el teorema de Bayes. Los estudios que comparan algoritmos de clasificación han encontrado que un clasificador bayesiano simple conocido como el clasificador Naive Bayes es comparable en rendimiento con el árbol de decisión y algunas redes neuronales. Los clasificadores bayesianos también han exhibido alta precisión y velocidad cuando se aplican a grandes bases de datos (Han, 2011).

2.1.4.11. K-means

El algoritmo k-means define el centroide de un grupo como el valor medio de los puntos dentro del grupo. Procede de la siguiente manera: primero, selecciona aleatoriamente k objetos del dataset, cada uno de los cuales inicialmente representa una media o centro del grupo. Para cada uno de los objetos restantes, se asigna un objeto al grupo al que es más similar, en función de la distancia euclidiana entre el objeto y la media del grupo. El algoritmo k-means luego mejora iterativamente la variación dentro del grupo. Para cada grupo, calcula la nueva media utilizando los objetos asignados al grupo en la iteración anterior; segundo, todos los objetos se reasignan utilizando las medias como nuevos centros de clúster. Las iteraciones continúan hasta que la asignación es estable, es decir, los grupos formados en la ronda actual son los mismos que los formados en la ronda anterior (Han, 2011).

2.1.4.12. Affinity Propagation

Crea grupos al enviar mensajes entre puntos de datos hasta la convergencia. A diferencia de los algoritmos de agrupamiento en clúster, como k-means, este algoritmo no requiere que el número de agrupamientos se determine o estime antes de ejecutar el algoritmo, para este propósito, los dos parámetros importantes son la preferencia, que controla la cantidad de ejemplares (o se utilizan prototipos), y el factor de amortiguamiento que frena la responsabilidad y la disponibilidad de los mensajes para evitar oscilaciones numéricas al actualizar estos mensajes.

2.1.4.13. Principal component analysis (PCA)

Supongamos que los datos a reducir consisten en tuplas o vectores de datos descritos por n atributos o dimensiones. El análisis de componentes principales (PCA) busca los k mejores vectores ortogonales $n - dimensionales$ que se pueden utilizar para representar los datos,

donde $k < n$. Los datos originales se proyectan así en un espacio mucho más pequeño, lo que resulta en una reducción de la dimensionalidad.

2.2. Revisión Literaria

En esta sección se describirán trabajos relacionados con el tema de la minería de datos en el campo educativo, se presenta su importancia, los métodos mas usados y los patrones que han resultado de las investigaciones realizadas.

2.2.1. Trabajos relacionados

En Eckert y Suénaga (2015) se analiza información académica con el objetivo de identificar factores que influyen sobre la deserción de los estudiantes , mediante la aplicación de minería de datos. La fuente de datos usada incluye información proporcionada al momento del ingreso de los estudiantes a la universidad y la que se genera durante el periodo de estudios, se aplicaron algoritmos de clasificación como árboles de decisión, redes bayesianas y reglas, se identificaron atributos que se relacionan fuertemente con la deserción y permanencia como: la cantidad de asignaturas aprobadas del primer año, el número de asignaturas cursadas, la edad de ingreso, la procedencia. Se realizo un proceso de clasificación con los atributos descritos anteriormente y se obtuvo un porcentaje de aciertos del 76 %.

Pereira (2010) tiene como objetivo determinar perfiles de bajo rendimiento académico y deserción estudiantil, para ello se usaron datos históricos académicos de estudiantes de pre-grado en la Universidad de Nariño. Para generar las reglas de clasificación su utilizó el algoritmo C4.5 y para las reglas de asociación, el algoritmo EquipAsso, basados en atributos como ingresos, edad, edad ingreso, valor de la matrícula, cantidad de materias perdidas y promedio acumulado.

La clases que se pretenden predecir son el rendimiento académico y la deserción, además de encontrar reglas de asociación entre los atributos descritos anteriormente.

En (Spositto y cols., 2010) se aplicó un proceso de descubrimiento del conocimiento en sobre los datos de estudiantes de la Universidad Nacional de La Matanza con el fin de encontrar un clasificador del rendimiento académico y detectar los patrones determinantes de la deserción estudiantil. El algoritmo FT (árboles funcionales, en inglés, functional trees) mostró un 78 % de instancias clasificadas correctamente en relación al rendimiento académico, en cuanto a la deserción el algoritmo J48 clasifico el 72 % de los registros correctamente. DE la evaluación a los generados tomando en cuentas su comprensibilidad, se encontró que el algoritmo J48 generó un árbol de decisión muy grande y por lo tanto poco comprensible y difícil de interpretar y que el árbol generado por el algoritmo FT no permite explicar el rendimiento académico y las causas de la deserción estudiantil.

2.2.2. Importancia y necesidad

Cuando un estudiante deserta de su carrera universitaria, las instituciones tienen una perdida cuantiosa de dinero, que puede llegar hasta el 29 % de los ingresos anuales por concepto de aranceles de matrícula según Torres y cols. (2016), es por esto que los procesos de extracción de conocimiento, con el fin de aplacar este fenómeno son de gran importancia para las universidades. El objetivo es detectar con anterioridad cuales son los estudiantes que presentan características relacionadas con la posibilidad de abandono y así proveer contención o ayuda especial (Eckert y Suénaga, 2015). El uso de un sistema académico de consejería puede ayudar a las autoridades en las actividades de toma de decisiones a través de disposiciones de gráficos visuales y diagramas de estudiantes que tienen dificultades en sus estudios (Sa y cols., 2014).El producto final de los modelos beneficia a estudiantes, docentes, padres y autoridades

académicas, no sólo para informar sobre la situación de los estudiantes cuyo desempeño podría estar asociado con una característica particular (positiva o negativa), sino también como asesoramiento para la toma de decisiones.La toma de decisiones efectivas depende de la rapidez con que se identifica y analiza información importante. Esta última afirmación resulta difícil de cumplir si se utilizan métodos clásicos de procesamiento y por tanto resulta necesario aplicar nuevas técnicas adaptadas específicamente para cada caso, que permitan identificar y encontrar información útil oculta en grandes bases de datos (Eckert y Suénaga, 2015).

2.2.3. Métodos mas usados

Las técnicas mas populares según el análisis realizado por (Mohamad y Tasir, 2013), en el ámbito educativo son: clustering, clasificación, patrones secuenciales, predicción y reglas de asociación. Identificaron de igual manera una tendencia presente hasta el año 2005 donde se venia aumentando el uso de reglas de asociación, puesto que requiere menos experiencia para comprenderlo, comparado con otros métodos. A partir del año 2005 los investigadores pasaron a usar frecuentemente técnicas de clasificación y clustering en los análisis. Debido a la gran cantidad de datos y variables que se deseaba incluir en los modelos, las reglas de asociación producían resultados complejos, que eran difíciles de comprender para los no expertos en minería de datos. Dentro del análisis se incluye además una recomendación para futuros investigadores en este campo, siempre pueden hacer comparaciones con diferentes algoritmos para el mismo conjunto de datos, y esto definitivamente sería algo para analizar, si se lograrán resultados similares utilizando algoritmos diferentes.

2.2.4. Patrones descubiertos

En múltiples trabajos (Pereira, 2010; Eckert y Suénaga, 2015; Pereira y cols., 2013), se encontró que el primer año de la carrera es donde adquieren mayor importancia las acciones de acompañamiento, puesto que en este periodo de tiempo se presenta alto riesgo de deserción. En Vera y cols. (2012) se observo que la variable que mas influye en la deserción es el nivel, se muestra que los estudiantes que se encuentran en primero, segundo, tercero, cuarto y quinto nivel son aquellos que tienen mayor tendencia a desertar en niveles como sexto y séptimo. En cuanto a la interacción con entornos virtuales de aprendizaje en Ordoñez Briceño (2013) se pudo constatar que la interacción del estudiante con estos entornos no posee una alta influencia para que los estudiantes deserten la carrera, puesto que existen estudiantes que han obtenido un nivel alto y medio en la interacción, sin embargo, han reprobado la asignatura y constan como desertores. Por lo tanto la dedicación y el desempeño que aplique el estudiante en las evaluaciones, son las que influirán en mayor porcentaje en la deserción de un estudiante. Según los resultados de Timarán y cols. (2013) los factores académicos que inciden en la deserción estudiantil, además de un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera, son la facultad a la que pertenece el estudiante y el área a la que pertenece las materias perdidas.

Capítulo III

METODOLOGÍA

3.1. Descripción del dominio de aplicación

En la presente sección se describe los objetivos que persigue la Universidad de Cuenca en el contexto del rendimiento académico, además de los datos que existen en los sistemas de información y finalmente los objetivos de minería de datos.

3.1.1. Objetivos del negocio

Los objetivos de la institución vinculados al presente trabajo fueron presentados en el Capitulo I, se detalla en términos generales que se busca generar una gestión eficiente, eficaz y efectiva en todas las actividades que desempeña la institución, encaminada en la búsqueda constante y sistemática de la excelencia académica y administrativa. (Universidad de Cuenca, 2017). En cuanto a un objetivo especifico se tiene el incremento de la tasa de titulación de grado y la tasa de retención inicial, además de lograr la acreditación institucional de carreras, en la que se incluye métricas relacionadas a la deserción y rendimiento académico.

3.1.2. Datos disponibles

Como paso inicial, se identificó en las bases de datos de los sistemas de información disponibles en la Universidad de Cuenca, los atributos relacionados a los objetivos y preguntas planteadas al inicio del trabajo. Por una parte, se tienen los datos académicos que contienen registros de calificaciones, asignaturas cursadas y asignaturas ofertadas de todos los estudiantes matriculados desde el año 2008 hasta el periodo actual, por otra existen los datos de la ficha socio-económica que es llenada todos los periodos académicos por los estudiantes.

3.1.3. Objetivos de la minería de datos

- Predecir si un estudiante egresa o abandona su carrera, basado en atributos históricos académicos y socio-económicos relevantes.
- Predecir si un estudiante va a perder al menos una asignatura, basado en su rendimiento académico pasado.
- Predecir si un estudiante aprueba o reprueba asignaturas especificas, al momento de su matriculación, basado en el datos académicos históricos de estudiantes similares.

3.2. Comprensión de los datos

A continuación se realiza una exploración de los datos almacenados por los sistemas de la universidad y se describen las tablas y campos mas importantes.

3.2.1. Descripción de los datos

Desde las bases de datos de los sistemas de información académica, se obtuvo un resumen de registros mediante consultas SQL [1], que pueden dar una idea de la dimensión del conjunto de datos a usar, en la Tabla 1 se presentan los casos de uso mas significativos (con datos desde al año 2008) para los objetivos de la minería de datos y el número de registros en las bases de datos. Se destaca el número de estructuras curriculares por estudiante puesto que ese será el máximo número de registros para los procesamientos posteriores. Estos casos son los que se pueden obtener sin realizar un procesamiento adicional, para obtener datos mas significativos es necesario realizar transformaciones sobre los datos, que se verán mas adelante.

Caso de Uso	Número Registros
Asignaturas Cursadas	1749806
Asignaturas No cursadas	1679247
Asignaturas Aprobadas	3799723
Mallas curriculares por estudiante	83371
matrículas en pre-grado (desde 2008)	62015
Número de estructuras curriculares	411
Número de estudiantes	82575
Número de grupos ofertados	80706

Tabla 1: Número de registros almacenados en base de datos

Las datos del Sistema de Gestión Académica, están divididos en varias bases de datos de acuerdo a la función que cumplen, por ejemplo se tiene una base para información relacionada a matrículas y otra para información de ofertas académicas, a continuación se describen las tablas que influyen en los objetivos de minería de datos del presente trabajo, agrupadas en las diferentes bases que tiene el sistema.

En la Tabla 2 se pueden observar las tablas del esquema de matrícula: *registro_persona*,

[1] por sus siglas en inglés Structured Query Language; en español lenguaje de consulta estructurada.

que contiene datos de la matrícula de cada estudiante por periodo de estudios; así mismo la tabla *registro_persona_asignatura* que contiene los datos de las matrículas de un estudiante por cada asignatura. En el esquema de matrícula de igual manera se tiene la tabla *detalle_registro_academico* que contiene información sobre las calificaciones y estado de aprobación de cada asignatura cursada por un estudiante, de la misma forma relacionada con las asignaturas cursadas esta la tabla *registro_academico* que contiene la información de las asignaturas aprobadas de la estructura curricular del estudiante; finalmente existe la tabla *registro_academico_por_detalle_registro_academico* que contiene la relación que existe entre una asignatura que el estudiante curso y la que corresponde a su estructura curricular. Los campos relevantes de estas tablas están descritos en la Tabla 3.

registro_persona	
Campo	Descripción
id_registro_persona	Llave primaria
id_periodo	Periodo lectivo de matrícula
id_persona	Identificador de estudiante
id_servicio	Identificador de estructura curricular
registro_persona_asignatura	
id_registro_persona_asignatura	Llave primaria
id_servicio_asignatura	Identificador único de una asignatura
id_grupo	Identificador único de un grupo dentro de una asignatura
veces_cursa	Indica que número de matrícula esta cursando un estudiante

Tabla 2: Tablas relacionadas a la matrícula del estudiante

En el esquema de servicios es donde se almacenan todos los conceptos relacionados a los servicios académicos que oferta la Universidad de Cuenca, ahí se tiene la tabla *servicio_asignatura* que contiene información especifica sobre asignaturas existentes en la universidad, de igual manera la tabla *asignatura_por_nivel* que contiene la relación entre una asignatura

detalle_registro_académico	
id_detalle_registro_académico	Lave primaria.
id _forma_aprobación	Indica si homologo, curso o aprobó por examen.
id_alumno	Identificador único de estudiante
valor	Nota asentada por el docente
id_estado_aprobación	Indica si el estudiante aprobó o reprobó
id_registro_persona_asignatura	Contiene la relación con la matrícula del estudiante.
registro_académico	
id_registro_académico	Llave primaria.
id_forma_aprobación	Indica si homologo, curso o aprobó por examen.
id_asignatura_por_nivel	Contiene la relación con la asignatura de su malla.
id_servicio	Contiene la relación con la asignatura.
nota	Nota con la que aprobó.
número_matrícula	Indica en que número de matrícula aprobó.
id_estado_aprobación	Indica si esta cursando, aprobó o no curso la asignatura.
id_estructura_curricular_x_estudiante	Indica que estructura contiene a la asignatura aprobada.
registro_académico_por_detalle_registro_académico	
id_detalle_registro_académico	Identificador de la asignatura cursada.
id_registro_académico	Identificador de la asignatura de la malla aprobada.
id_forma_aprobación	Indica si homologo, cursó o aprobó por examen.

Tabla 3: Tablas relacionadas a la aprobación de asignaturas

y el nivel de la malla en la que se encuentra, también entre asignaturas se pueden establecer equivalencias que están descritas en la tabla *equivalencia_asignatura*. Los campos específicos de están tablas se describen en la Tabla 4. Cabe destacar que existe la tabla *servicio* que engloba todos los conceptos y que contiene información de: estructuras curriculares, asignaturas y niveles.

El esquema en el que se tienen la información de asignaturas disponibles para la matriculación de estudiantes por periodo de estudio es oferta, este incluye la tabla con el mismo nombre que contiene la información del conjunto de asignaturas ofertadas que las facultades ponen a

equivalencia_asignatura	
Campo	Descripción
id_equivalencia_asignatura	Llave primaria
id_servicio_asignatura	Indica la asignatura principal
id_servicio_asignatura_equivalente	Indica la asignatura equivalente
asignatura_x_nivel	
id_asignatura_x_nivel	Llave primaria
id_servicio_asignatura	Identificador de la asignatura
id_nivel	Indica en que nivel se encuentra la asignatura
servicio_asignatura	
id_servicio_asignatura	Llave primaria
créditos	Contiene el número de créditos.
horas	Contiene el número de horas, si es de rediseño

Tabla 4: Tablas relacionadas a las asignaturas de la universidad

disposición de los estudiantes, para saber específicamente los cupos y carrera de la asignaturas se debe ir a la tabla *asignatura_por_oferta*, si se necesita información sobre horarios y docentes se puede encontrar en la tabla *grupo*. Los campos mas relevantes de cada tabla se presenta en la Tabla 5.

Para extraer los datos de la ficha socio-económico fue necesario acceder a la base de datos del sistema antiguo de la Universidad de Cuenca (eSIUC), debido a que al momento de realizar el trabajo no se tenían estos datos en el nuevo sistema. Los campos de la ficha se indican en la Tabla 6

Antes de comenzar los experimentos se considero necesario establecer un periodo base para realizar las predicciones y experimentos, para esto se analizó la diferencia entre el número de matrículas por cada par de periodos consecutivos de las personas que abandonaron la carrera. El resultado se puede ver en la Figura 1, ahí se puede apreciar que los periodos en los cuales los

oferta	
Campo	Descripción
id_oferta	Llave primaria
id_servicio	Contiene la estructura curricular que se oferta
nombre	Contiene el nombre de la oferta
id_periodo	Indica en que periodo se oferta
asignatura_por_oferta	
id_asignatura_por_oferta	Llave primaria
id_servicio_asignatura	Identificador de asignatura ofertada
id_unidad_oferta	Contiene la unidad que realiza la oferta
id_periodo	Contiene el periodo en que se oferta
créditos	número de créditos
grupo	
id_grupo	Lave primaria
id_asignatura_por_oferta	Contiene la asignatura a la que pertenece el grupo
id_docente	Indica el docente principal del grupo

Tabla 5: Tablas relacionadas a oferta de asignaturas

estudiantes abandonan la carrera son el tercero y quinto. Debido que para realizar la predicción del rendimiento en el tercer periodo se tendrían datos de solo dos ciclos anteriores, se escogió realizar el análisis para el quinto periodo de estudios, basado en las calificaciones de los cuatro primeros ciclos. Seria interesante en trabajos posteriores realizar el análisis para cada periodo de estudio en busca de modelos específicos.

sse_alumnos	
Campo	Descripción
CEDULA	Identificación del estudiante
PERLEC_ID	Identificador del periodo de estudios
FECHA_NACIMIENTO	Fecha de nacimiento del estudiante
TENVIV_ID	Indica si la vivienda es propia, arrendada o cedida
MATVIV_ID	Indica si la vivienda es de ladrillo o adobe
ZONA_VIVIENDA	Indica si la vivienda esta ubicada en la zona urbana o rural
TIPVIV_ID	Indica si la vivienda es una casa, villa, media agua, etc..
ARRIENDO	Indica el valor del arriendo si lo paga.
NRO_PROPIEDADES_RENTERAS	Indica si se tienen propiedades en arriendo
TVCABLE	Indica si tiene o no servicio de televisión por cable
NRO_VEHICULO	Indica el número de vehículos que posee
AVALUO_VEHICULOS	Indica el avalúo total de vehículos
NUMINTEGRATES	Indica el número de integrantes del núcleo familiar
NUMESTUDIANTRES	Indica el número de familiares que estudian
NUMHIJOS_MEN6	Indica el número de hijo menores de 6 años
DESCRIPCION	Indica la ocupación del jefe de familia
ROLFAM_ID	Indica quien es el jefe de familia
PROVINCIA	Provincia de nacimiento
CANTON	Cantón de nacimiento
PAIS	País de nacimiento
INGRESOS_FAMILIARES	Cantidad de dinero que la familia recibe mensualmente
EGRESOS_FAMILIARES	Cantidad de dinero que la familia gasta mensualmente

Tabla 6: Campos de la ficha socio-económica

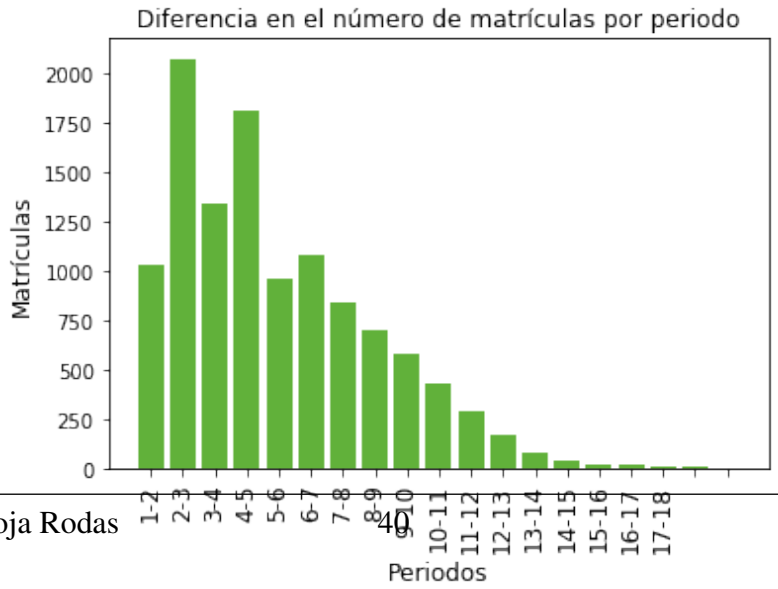


Figura 1: Diferencia de número de matrículas entre periodos para desertores

A partir de la siguiente etapa del proceso, se procedió a dividir el trabajo entre los tres objetivos de minería de datos, es decir, se va a realizar un experimento por cada objetivo y cada experimento tiene diferentes etapas de preparación de datos, construcción de modelo y evaluación que se describen a continuación:

3.3. Experimento 1 : Predicción de deserción basado en atributos académicos y socio-económicos relevantes

En esta sección se presenta el proceso de la construcción de un modelo predicativo, se parte con l preparación de los datos de entrada, luego se describe el proceso de transformación de los datos y finalmente la construcción y evaluación del modelo generado.

3.3.1. Preparación de los datos

Aquí se describe el proceso para la extracción de los datos que incluye: limpieza, filtrado y finalmente un análisis de las variables que mas influyen en la deserción. El proceso ETL realizado para la transformación de datos se indica en el Anexo **??**.

3.3.1.1. Datos para el análisis

Para el primer objetivo "Predecir si un estudiante egresa o abandona su carrera, basado en atributos históricos académicos y socio-económicos relevantes"se decidió utilizar los datos académicos de los primeros dos años de estudio y los datos de la ficha socio-económica del ultimo periodo cursado.

3.3.1.2. Transformación de los datos

En esta etapa se obtuvo la variable de respuesta para el primer experimento (completo), que toma un valor de 0 si el estudiante abandonó sus estudios y de 1 si es que completo todas las asignaturas de su malla curricular. Para determinar si el estudiante abandonó sus estudios se tomó como referencia que no esté matriculado en los últimos tres periodos de estudios en ninguna asignatura. A continuación, se procedió a calcular el promedio por periodo de estudios de cada estudiante agrupando las notas de las asignaturas aprobadas en el registro académico, aquí se tomó en cuenta la variable del número de matrícula para tomar solo el 50 % o 33 % de la nota si ésta fue aprobada en segunda o tercera matrícula respectivamente. Esto fue necesario debido a que no se tiene una relación establecida entre las asignaturas reprobadas y el registro académico de los estudiantes, y por lo tanto no se pudo realizar un promedio con las asignaturas reprobadas que hubiese sido lo óptimo. En esta etapa se calculó también el número de asignaturas que tomó el estudiante, puesto que la carga horaria puede ser un factor determinante para que un estudiante abandone su carrera. A partir de la obtención de estos campos, fue necesario pasarlos a un sólo registro por estudiante y estructura curricular, puesto que se tenia un registro por cada periodo para cada estudiante. Al finalizar este proceso se obtuvieron 43313 registros, de los cuales 12790 corresponden a estudiantes que completaron su malla curricular y 30523 a estudiantes que aún tienen asignaturas pendientes; de este último grupo 13546 corresponden a estudiantes que se retiraron y 16977 que aun están cursando sus estudios. A partir de estos registros, se tomó a los estudiantes que cursaron al menos 4 ciclos de estudios para realizar el análisis, con esta condición quedan 9987 estudiantes que completaron su malla y 6588 que abandonaron sus estudios. Como paso final se agregaron los datos de la última ficha socio-económica y el promedio de ingresos (ingresos_prom) y egresos (egresos_prom) de los 4 periodos de estudios a los registros de cada alumno, para tener así un solo registro con los datos

académicos y socio-económicos, en la Tabla 7 se describen los campos con los que se realizara la selección de atributos que influyen el la deserción, se destaca que los campos calculados fueron: el promedio de calificaciones por cada periodo, el número de asignaturas cursadas por cada periodo, el promedio de ingresos económicos, el promedio de egresos económicos y la variable de respuesta egresó o abandonó.

Conjunto de datos, experimento 1	
Campo	Descripción
Completo	Variable de respuesta (egresó = 1 , desertó = 0)
prom_1	Promedio en el primer ciclo de estudios
prom_2	Promedio en el segundo ciclo de estudios
prom_3	Promedio en el tercer ciclo de estudios
prom_4	Promedio en el cuarto ciclo de estudios
mat_1	número asignaturas cursadas en el primer ciclo de estudios
mat_2	número asignaturas cursadas en el segundo ciclo de estudios
mat_3	número asignaturas cursadas en el tercer ciclo de estudios
mat_4	número asignaturas cursadas en el cuarto ciclo de estudios
ingresos_prom	promedio de ingresos económicos de los cuatro periodos
egresos_prom	promedio de ingresos económicos de los cuatro periodos
TENVIV_ID	Indica si la vivienda es propia, arrendada o cedida
MATVIV_ID	Indica si la vivienda es de ladrillo o adobe
ZONA_VIVIENDA	Indica si la vivienda esta ubicada en la zona urbana o rural
TIPVIV_ID	Indica si la vivienda es una casa, villa, media agua, etc..
ARRIENDO	Indica el valor del arriendo si lo paga.
NRO_PROPIEDADES_RENTERAS	Indica si se tienen propiedades en arriendo
TVCABLE	Indica si tiene o no servicio de televisión por cable
NRO_VEHICULO	Indica el número de vehículos que posee
AVALUO_VEHICULOS	Indica el avalúo total de vehículos
NUMINTEGRATES	Indica el número de integrantes del núcleo familiar
NUMESTUDIANTRES	Indica el número de familiares que estudian
NUMHIJOS_MEN6	Indica el número de hijo menores de 6 años
PROVINCIA	Provincia de nacimiento
CANTON	Cantón de nacimiento
PAIS	País de nacimiento

Tabla 7: Datos de entrada para el modelado

3.3.1.3. Limpieza y filtrado de datos

Antes de comenzar con el análisis es necesario codificar, mediante una transformación, los datos categóricos existentes en el conjunto de datos a variables numéricas, estos son: PROVINCIA, CANTON, PAIS, TENENCIA DE VIVIENDA, MATERIAL DE VIVIENDA, ZONA DE VIVIENDA Y TIPO DE VIVIENDA. A continuación se aplicó un proceso para quitar los valores atípicos (outlier), principalmente en los campos de ARRIENDO y número DE INTEGRANTES donde se identificó mediante un análisis de quartiles que habían datos que posiblemente estaban fuera de los rangos normales; con el fin de evitar errores en el procesamiento de los algoritmos, se llenaron con 0 los registros que tenían valores nulos para los campos de *ingresos_prom*, *egresos_prom* y *NUMHIJOS_MEN6*.

En esta etapa se analizó la cantidad de registros agrupados por la variable de respuesta, aquí se encontró que las clases no estaban balanceadas, el 40 % de registros correspondían a la clase 0 (desertó) y el 60 % a la clase 1 (egresó), entonces se aplicó un proceso de sub-muestreo para equilibrar los registros de las dos clases, primero se aplicó el algoritmo Tomek links (Tomek, 1976) , el mismo que remueve registros de clase mayoritaria que sean similares a la clase minoritaria; luego se aplicó un proceso de sub-muestreo aleatorio para que las dos clases queden con el mismo número de registros.

3.3.1.4. Selección de datos para el análisis

Los datos seleccionados para la construcción del modelo del primer experimento son: el número de materias cursadas por periodo, el promedio de notas por periodo y los datos de la ficha socio-económica. Una vez que se tiene seleccionado los datos iniciales se procedió a realizar un análisis de atributos, con el objetivo de reducir el número de atributos a usar, reducir el sobre-ajuste y mejorar la generalización de los modelos resultantes. Además es importante

este análisis para tener una visión general de los atributos y conocer como se relacionan con la variable de respuesta.

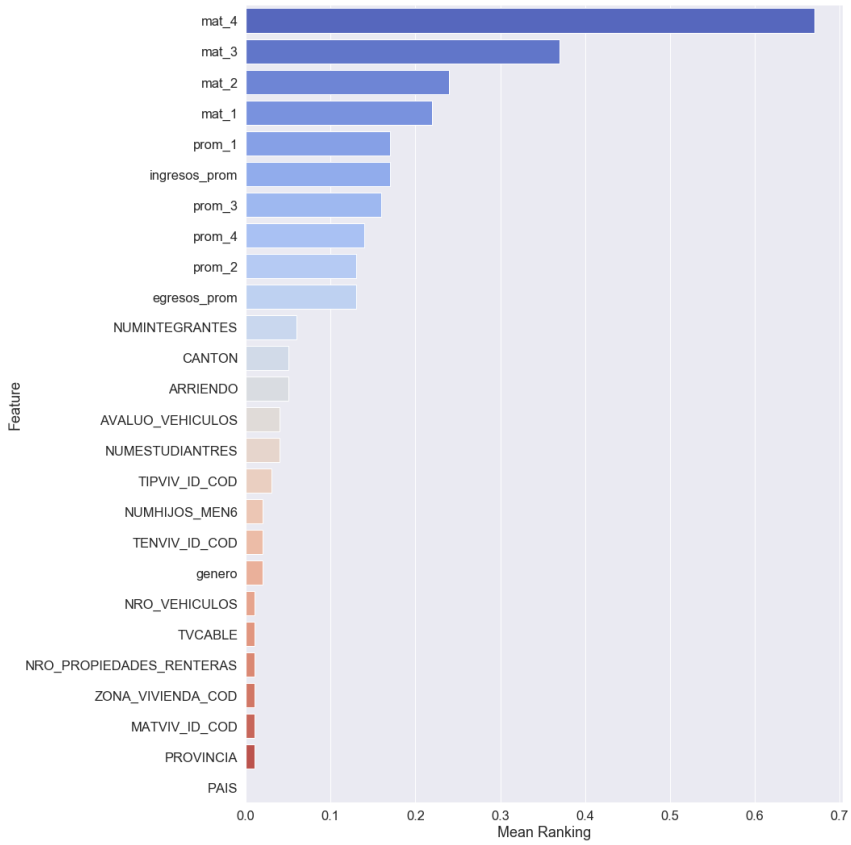


Figura 2: Ranking de promedio de variables

Para este análisis se utilizaron tres métodos: Stability Selection via Randomized Lasso Method, Linear Model Feature Coefficients y Random Forest Feature Selection; se combinaron los resultados de cada uno mediante el cálculo del promedio y se formo un ranking de atributos. En la Figura 2 se puede observar que los atributos que mas influyen en la variable de respuesta son: los promedios de los cuatro primeros ciclos, el número de asignaturas matriculadas en los cuatro primeros periodos, el promedio de ingresos económicos, el promedio de egresos económicos. Se podría concluir que los datos mas importantes para la construcción del modelo son

Conjunto de datos, experimento 1	
Campo	Descripción
Completo	Variable de respuesta (egresó = 1 , desertó = 0)
prom_1	Promedio en el primer ciclo de estudios
prom_2	Promedio en el segundo ciclo de estudios
prom_3	Promedio en el tercer ciclo de estudios
prom_4	Promedio en el cuarto ciclo de estudios
mat_1	número asignaturas cursadas en el primer ciclo de estudios
mat_2	número asignaturas cursadas en el segundo ciclo de estudios
mat_3	número asignaturas cursadas en el tercer ciclo de estudios
mat_4	número asignaturas cursadas en el cuarto ciclo de estudios
ingresos_prom	promedio de ingresos económicos de los cuatro periodos
egresos_prom	promedio de ingresos económicos de los cuatro periodos

Tabla 8: Atributos seleccionados para el modelado

los relacionados al ámbito académico, los ingresos y egresos familiares; en cambio las variables restantes que tienen que ver con el aspecto socio-económico no son decisivas para saber si un estudiante egresa o no.

Con los atributos escogidos se procedió a normalizar los datos para tener una escala constante entre el promedio de calificaciones y el número de asignaturas tomadas por ciclo, esto permite que el peso de cada atributo usado sea el mismo al realizar la construcción del modelo. Los campos del conjunto de datos (dataset) final a usar se describen en la Tabla 8.

3.3.2. Modelado

En esta sección se describe el proceso para la construcción del modelo, las técnicas usadas y su evaluación. El código usado para la construcción del modelo se indica en el Anexo **??**

3.3.2.1. Técnicas usadas

Para la construcción del modelo se empleo 8 técnicas, en base a una clasificación inicial, se identificó a los tres algoritmos con mejor precisión, y luego se realizó un proceso de optimización de parámetros para cada uno y finalmente se obtuvo el algoritmo de mejor desempeño. Los 8 algoritmos iniciales son:

1. K vecinos más cercanos (en inglés, k-nearest neighbors)
2. Clasificación de vectores de soporte (en inglés, Support Vector Classification)
3. Árbol de decisión
4. Bosques aleatorios
5. Perceptron multicapa
6. Naive Bayes
7. Ada Boost
8. Análisis cuadrático discriminante (en inglés, Quadratic Discriminant Analysis)

3.3.2.2. Generación de pruebas de modelo

Para la validación del modelo se uso la métrica de exactitud (accuracy), que es la relación entre el número total de muestras y el número de aciertos del clasificador; en cuanto a los datos a usar en la evaluación, se realizó mediante el método de validación cruzada con 10 subconjuntos de datos (folds). Al final la métrica se obtuvo del promedio de los 10 subconjuntos.

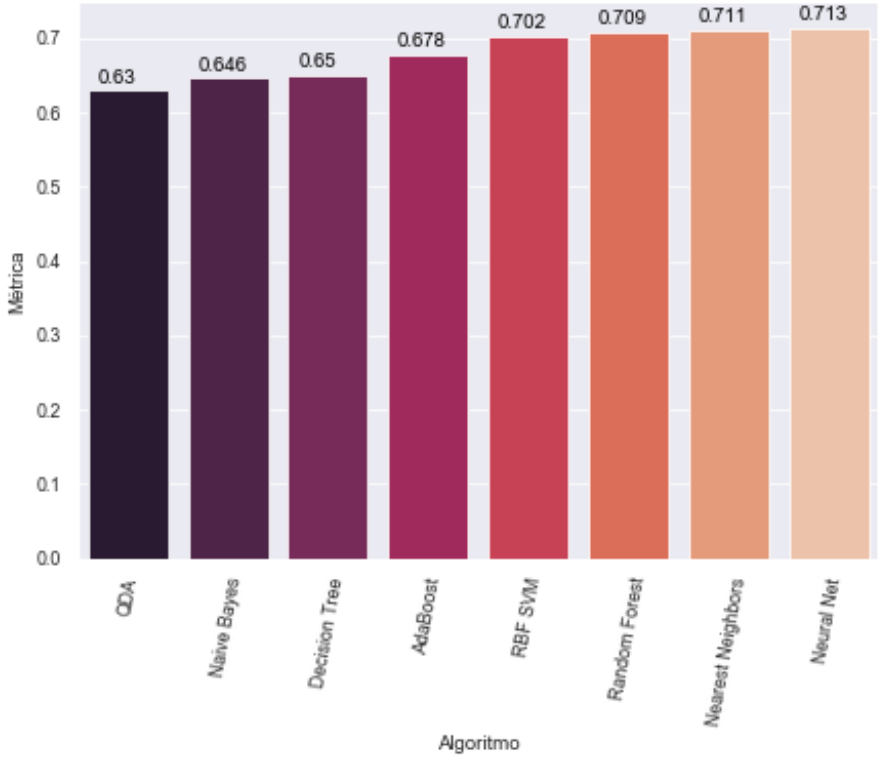


Figura 3: Exactitud de los 8 algoritmos iniciales

3.3.2.3. Construcción del modelo

Para realizar la construcción del modelo se uso la librería de sklearn[2], disponible dentro de Python. Como primer paso se ejecutó la clasificación con los 8 algoritmos iniciales, para esto se tomó el 80 % de los datos para entrenamiento y el 20 % restante para validación. Como resultado se obtuvo los valores de exactitud descritos en la Figura 3, donde se puede notar que los tres algoritmos de mejor desempeño fueron: bosques aleatorios, k-vecinos mas cercanos y perceptron multicapa.

[2]Scikit-learn es una biblioteca para aprendizaje de máquina de software libre para el lenguaje de programación Python

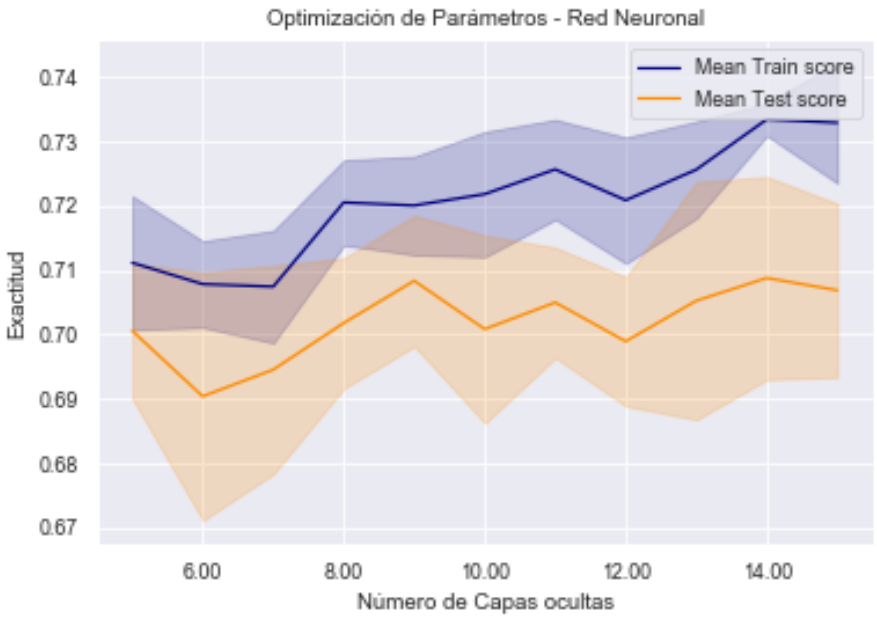


Figura 4: Influencia del número de capas en la exactitud del modelo

Con estos resultados se pasó a la siguiente etapa, en la cual se ejecutó un proceso de optimización de parámetros para los tres algoritmos antes mencionados, con el objetivo de obtener el mejor modelo posible. Este proceso toma como entrada una matriz de valores de los diferentes parámetros que acepta cada algoritmo, con esto realiza la clasificación con cada combinación de parámetros posibles, y evalúa los modelos resultantes con validación cruzada de 3 subconjuntos, al final se obtienen los parámetros que alcanzaron la mejor exactitud.

Se inició con el algoritmo de perceptron multicapa, en el cual se especificaron los siguientes parámetros: número máximo de iteraciones (150, 300, 500, 1000), alfa (1.e-01, 1.e-02, 1.e-03, 1.e-04, 1.e-05, 1.e-06) y el número de neuronas en la capa oculta (5, 6, 7, 8, 9, 10, 11, 13, 14). En total el proceso corre 720 instancias del modelo de los cuales como resultado se obtuvo que los parámetros que alcanzan el mejor puntaje son : alfa = 0.001, número de neuronas en la capas oculta = 14 y número máximo de iteraciones = 500, con una exactitud de 0.715. En la Figura 4

se puede observar como el número de capas ocultas influye en el desempeño del modelo, tanto para el puntaje de entrenamiento y validación; ahí se puede observar que el puntaje tiene una tendencia a aumentar en relación al número de neuronas, esta relación es mas marcada para el puntaje de entrenamiento, puesto que para el puntaje de validación se puede observar que llega al máximo en 14 neuronas.

MLPClassifier				
{'alpha': 0.001, 'hidden_layer_sizes': 14, 'max_iter': 500, 'solver': 'lbfgs'}				
	precision	recall	f1-score	support
N	0.70	0.68	0.69	1128
S	0.70	0.72	0.71	1192
Exactitud (accuracy) 80/20	**0.699**			
Exactitud promedio (accuracy) 10 fold	**0.709**			

Tabla 9: Resultado del algoritmo Perceptron multicapa con parámetros óptimos

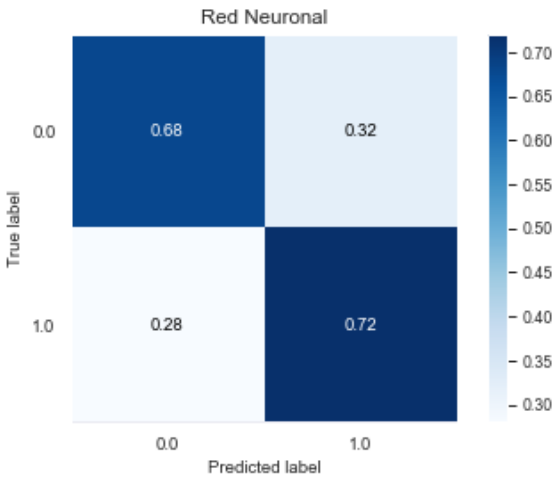


Figura 5: Matriz de confusión de algoritmo Perceptron multicapa

Una vez que se tiene los parámetros óptimos, se procede a aplicar el algoritmo de redes neuronales sobre el conjunto de datos completo mediante el método de evaluación de valida-

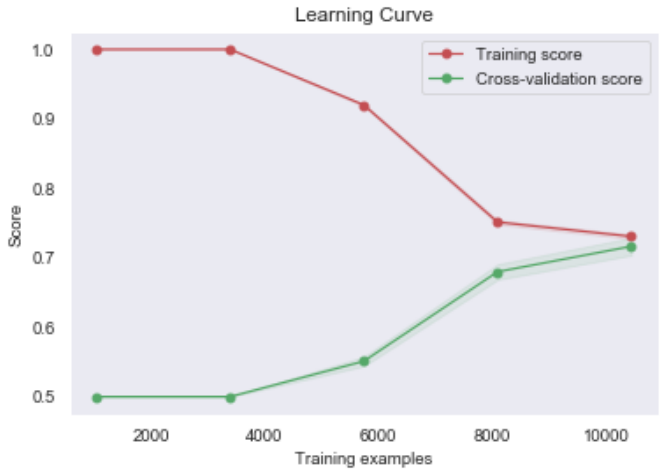


Figura 6: Curva de aprendizaje algoritmo Perceptron multicapa

ción cruzada y así obtener el puntaje final de la exactitud del modelo. En la Tabla 9 se puede observar un resumen de los resultados de la clasificación, además en la Figura 5 se puede observar la matriz de confusión del algoritmo. También se ha incluido como referencia a la curva de aprendizaje en la Figura 6 obtenida con el algoritmo, en la cual se puede observar que tiene el comportamiento deseado para un clasificador de buenas condiciones, es decir, que a medida que el número de muestras de entrenamiento va aumentando el puntaje de entrenamiento y validación convergen hacia un solo punto.

El siguiente algoritmo escogido para optimizar es k-vecinos mas cercanos, para el cual se especificaron los siguientes parámetros: número de vecinos con valores del 1 al 30 y el parámetro p, que especifica que función se usará para calcular la distancia entre miembros del cluster, para valores de: 1 usa la distancia de manhattan, 2 usa la distancia euclideana y para valores mayores se usa la distancia de minkowski, en el presente caso se usaron los valores del 1 al 4. En total el proceso corre 360 instancias de clasificación para cada combinación de parámetros y para cada uno de los 3 subconjuntos de datos, al finalizar se encontró que los mejores parámetros son: número de vecinos = 24 y p = 1, con una exactitud de 0.725. En

la Figura 7 se puede observar como a medida que el número de vecinos cercanos aumenta el puntaje del conjunto de datos de entrenamiento disminuye hasta estabilizarse alrededor del valor 20, en cambio para el conjunto de datos de validación se puede ver que mientras el valor del número de vecino aumenta, de igual manera lo hace el puntaje, hasta estabilizarse en valores mayores a 10. Se puede notar así mismo que la variabilidad entre el puntaje de los diferentes subconjuntos de datos es mínima, esto esta denotado por el área alrededor de la media del puntaje.

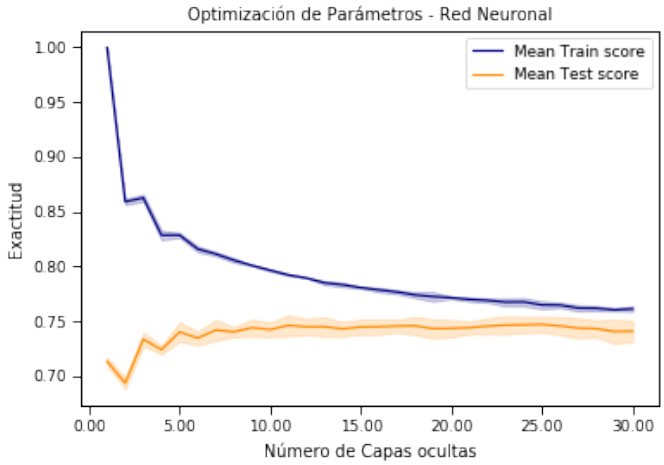


Figura 7: Influencia del número de vecinos en la exactitud del modelo

Con los parámetros óptimos identificados se ejecuto el algoritmo de k-vecinos más cercanos con todo el conjunto de datos, y se lo evaluó con el método de validación cruzada con 10 subconjuntos de datos, en la Tabla 10 se puede ver los resultados obtenidos, donde se destaca el valor de la exactitud del modelo (0.726); se puede notar que la clasificación tiene una buena exactitud para la clasificación de ambas clases, esto se comprueba mediante la matriz de confusión presentada en la Figura 8, donde se puede ver que el 74 % de registros que son de la clase 1 fueron clasificados correctamente, de igual manera el 71 % de registros de la clase 0.

Una forma de analizar la calidad del modelo es observar su curva de aprendizaje, en la cual

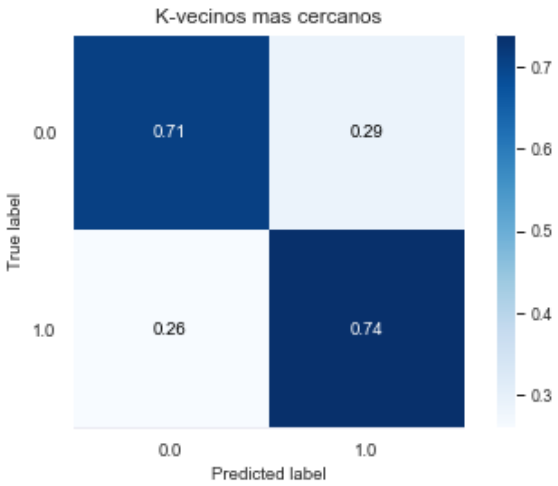


Figura 8: Matriz de confusión del algoritmo k-vecinos mas cercanos

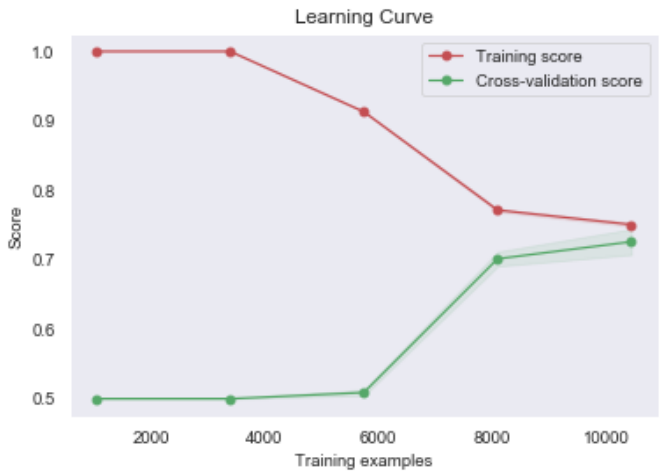


Figura 9: Curva de aprendizaje del algoritmo k-vecinos mas cercanos

KNeighborsClassifier				
{'n_neighbors': 24, 'p': 1}				
	precision	recall	f1-score	support
N	0.72	0.71	0.71	1128
S	0.73	0.74	0.75	1192
Exactitud (accuracy) 80/20	**0.724**			
Exactitud promedio (accuracy) 10 fold	**0.726**			

Tabla 10: Resultado del algoritmo k-vecinos mas cercanos con parámetros óptimos

se espera que el puntaje del conjunto de datos de validación siempre sea menor al de los datos de entrenamiento, que para conjuntos de datos pequeños el modelo exhiba sobre-ajuste y para conjuntos de datos grandes se observe sub-ajuste; entonces si se grafica el puntaje del conjunto de datos de entrenamiento y validación en función del número de muestras es deseable que los valores converjan hacia un punto medio. El modelo obtenido en el presente trabajo muestra el comportamiento descrito anteriormente, como se puede observar en la Figura 9.

Finalmente, se procedió a realizar el proceso de optimización de parámetros para el algoritmo de bosques aleatorios, se establecieron los parámetros de: número de arboles (5, 10, 15, 20, 25, 30, 35, 40), número máximo de atributos ('auto', 'sqrt', 'log2'), la profundidad máxima del árbol (5, 10, 15, 20, 25, 30, 35, 40) y el criterio para medir la calidad de la división ('gini','entropy'). El algoritmo realiza 1152 clasificaciones, una para cada combinación de parámetros y para los 3 subconjuntos de datos de validación cruzada, se obtuvo que los mejores parámetros son: número de arboles=40, número máximo de atributos='auto', profundidad máxima = 20 y criterio = 'entropy'; con una exactitud de 0.731. En la Figura 10 se puede observar como el parámetro de la profundidad máxima de cada árbol afecta el desempeño del modelo, tanto en el subconjunto de datos de entrenamiento, como en el de validación; se puede notar que el modelo esta teniendo sobre-ajuste para los datos de entrenamiento a medida que aumen-

ta la profundidad máxima, esto afecta a la generalización del modelo, es decir, su habilidad de predecir basado en datos nuevos.

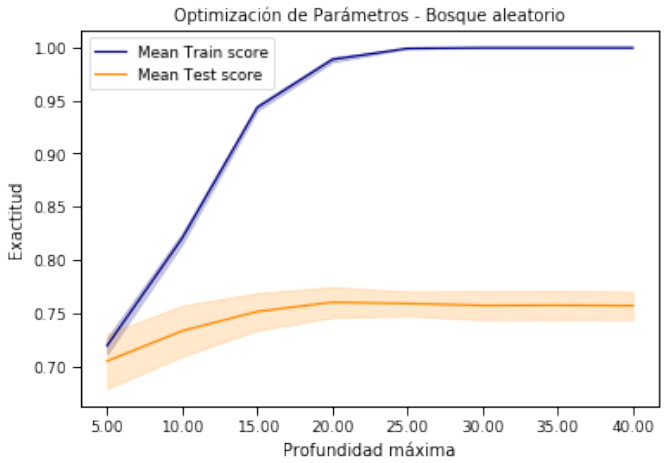
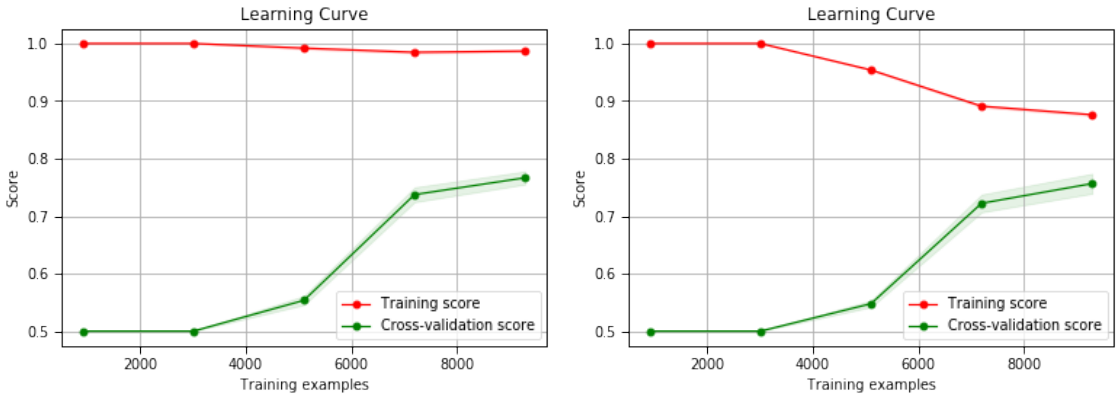


Figura 10: Influencia de la profundidad máxima en la exactitud del modelo

Usando validación cruzada, se procedió a evaluar el modelo con los parámetros óptimos; con esto se obtuvo que su exactitud es 0.737. Para verificar que el modelo no tenga sobre ajuste se procedió a graficar la curva de aprendizaje (Figura 11a), ahí se puede ver como el modelo para un número de muestras pequeño tiene exactitud alta para el subconjunto de entrenamiento, pero a medida que se aumenta el número de muestras este puntaje no disminuye, que sería lo esperado en un modelo de buenas condiciones, sino que se mantiene con un puntaje alto. Esto se puede dar debido a un número alto de árboles en el bosque aleatorio o también por la máxima profundidad establecida para cada árbol, para encontrar un modelo que no exhiba sobre-ajuste se procedió a realizar nuevamente el proceso de optimización de parámetros, esta vez con un número menor de árboles (5, 10, 15, 20, 25, 30) y profundidad máxima (1, 5, 7, 8 ,10, 12). Como parámetros óptimos se obtuvo: número de arboles = 30, profundidad máxima = 12, criterio = 'gini' y número de atributos = 'auto', con estos parámetros se ejecuto la clasificación nuevamente obteniendo los resultado descritos en la Tabla 11, se pudo notar que la variación

RandomForestClassifier				
{'criterion': 'gini', 'max_depth': 12, 'max_features': 'auto', 'n_estimators': 30}				
	precision	recall	f1-score	support
N	0.73	0.69	0.71	1128
S	0.72	0.76	0.74	1192
Exactitud (accuracy) 80/20	**0.726**			
Exactitud (accuracy) 10 fold	**0.733**			

Tabla 11: Resultado del algoritmo bosques aleatorios sin sobre-ajuste



(a) Con sobre-ajuste

(b) Sin sobre-ajuste

Figura 11: Curva de aprendizaje del algoritmo bosques aleatorios

en la exactitud no fue considerable entre los dos modelos (0.04), teniendo un mejor puntaje el modelo con sobre-ajuste. Lo relevante para este modelo es analizar el sobre-ajuste sobre los datos de entrenamiento, para esto se gráfico la curva de aprendizaje que se puede ver en la Figura 11b, aquí se puede notar que el puntaje de entrenamiento y validación convergen hacia un punto, el comportamiento deseable para un modelo de clasificación.

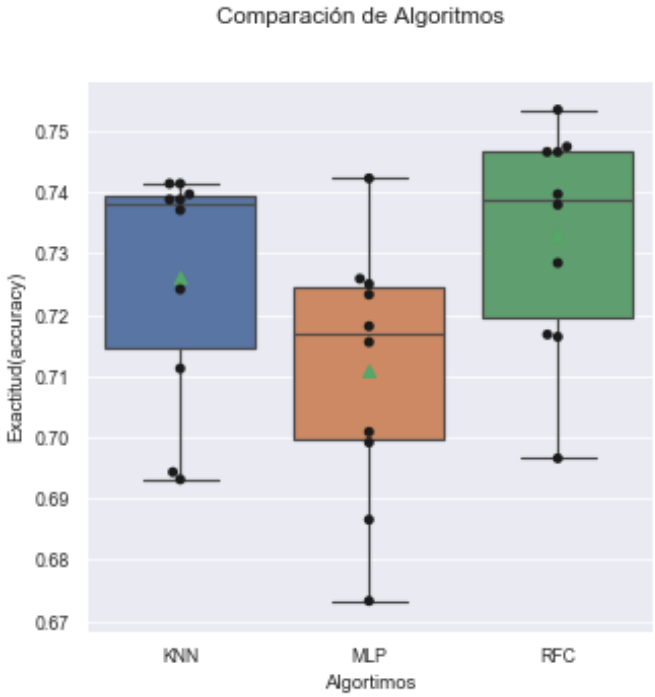


Figura 12: Distribución de puntajes de los 3 algoritmos

3.3.2.4. Evaluación del modelo

Finalmente se tiene los puntajes de los 3 algoritmos seleccionados para la construcción del modelo, estos fueron ejecutados sobre la totalidad de los datos usando el método de evaluación de validación cruzada con 10 subconjuntos, la exactitud media del algoritmo k-vecinos mas cercanos (KNN) fue 0.726, del algoritmo de perceptron multicapa(MLP) fue 0.711 y del bosques aleatorios(RFC) fue 0.732. En la Figura 12 se observa la distribución de los puntajes de cada algoritmo para los 10 subconjuntos de datos, en base a este gráfico se puede concluir que el algoritmo que produce un mejor modelo es el de bosques aleatorios (RFC) debido a que la media de los puntajes es la más alta.

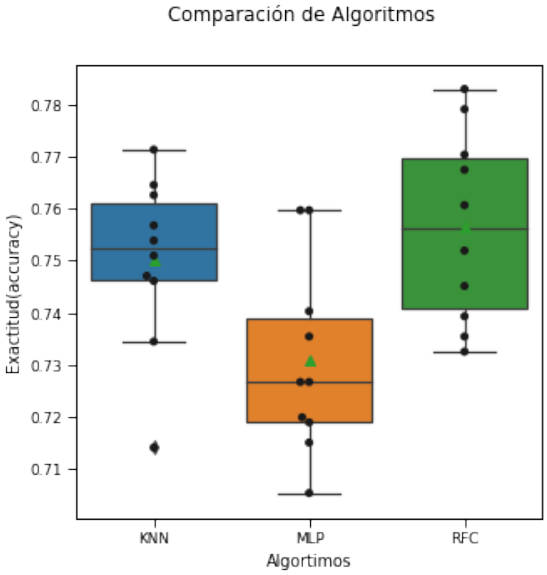


Figura 13: Distribución de puntajes de los 3 algoritmos con variables académicas

Cabe destacar que se realizó el mismo experimento tomando en cuenta solo las variables académicas, es decir, sin ingreso y egresos económicos; los resultados se presentan en la Figura 13. La exactitud media del algoritmo k-vecinos mas cercanos (KNN) fue 0.750, del algoritmo de perceptron multicapa(MLP) fue 0.731 y del bosques aleatorios(RFC) fue 0.756. Se puede ver que el algoritmo con mejor media es bosques aleatorios (RFC).

Se observa una leve mejoría (0.024) en los puntajes si se usan solo variables académicas.

3.4. Experimento 2 : Predicción de reprobación en al menos una asignatura, basado en rendimiento académico pasado.

En esta sección se presenta los procesos de transformación de datos y construcción del modelo para el segundo objetivo de minería de datos.

3.4.1. Preparación de los datos

A continuación se describen los procesos de limpieza y transformación de datos, además del calculo de atributos especiales. El proceso ETL realizado para la transformación de datos se indica en el Anexo **??**.

3.4.1.1. Datos para el análisis

Para el presente experimento, se utilizo los datos académicos de los dos primeros años de estudio, esto comprende, todas las calificaciones finales de asignaturas cursadas por los estudiantes, sin importar que hayan aprobado o reprobado. Con esto se pretende obtener las notas de los cuatro primeros ciclos y predecir si un estudiante va a reprobar al menos una asignatura en el quinto ciclo.

3.4.1.2. Transformación de los datos

En esta etapa, se calculó el promedio de calificaciones por periodo de estudio en una malla curricular, además se calculó el promedio asignaturas aprobadas así mismo por periodo de estudio, este campo se uso mas adelante para el calculó de la variable de respuesta. En este punto, en

los datos se tenían varios registros por estudiante, por esto fue necesario ejecutar un proceso de normalización de datos, de esta forma se obtuvo un registro por estudiante y estructura curricular, en el cual consta el promedio de los cuatro primeros periodos de estudio y el porcentaje de aprobación de asignaturas por periodo de los 5 primeros ciclos de estudio. En este punto se optó por realizar el proceso agrupando datos por estructura curricular, es decir ejecutar el proceso de construcción del modelo de predicción para cada carrera. Como base se tomo la carrera de Enfermería por contar con el mayor número de estudiantes en su malla vigente (991).

3.4.2. Limpieza y filtrado de los datos

Se filtraron los registros que tengan valor en el campo de la nota promedio para el quinto ciclo, de esta forma se descartó los estudiantes de ciclos inferiores. Luego de esto se identificaron los campos con valores nulos y se los reemplazó con 0, se asume que, si no tenían valores en esos campos no se matricularon en ninguna asignatura ese periodo.

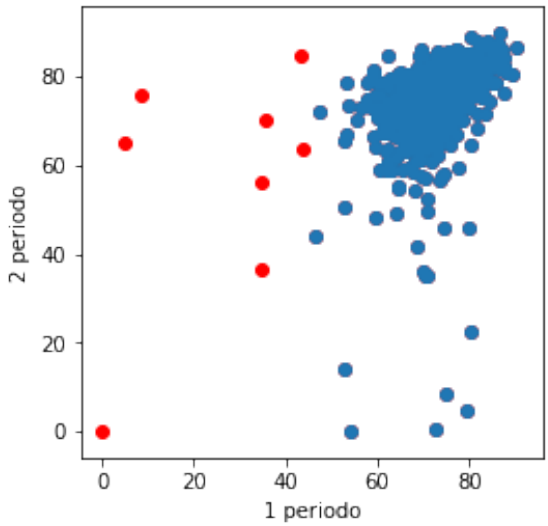


Figura 14: Outliers eliminados

Como siguiente paso se eliminaron los outliers de los campos de promedio de notas, para eliminar los datos considerados como ruido para el algoritmo, en el Gráfico 14 se puede apreciar en color rojo los outliers eliminados del campo promedio de notas del primer ciclo de estudios. Como resultado de este proceso se tiene 730 registros, 520 para la clase no reprueba y 210 para la clase reprueba. A partir de estos registros se ejecuto un proceso de balanceo de clases, mediante sub-muestreo aleatorio, quedando 210 registros para cada clase.

El siguiente paso de transformación de datos, se aplicó un proceso de reducción de dimensiones del conjunto de datos, usando el método Principal Component Analysis (PCA), este extrae componentes que son una combinación lineal de las variables originales, en el caso de este trabajo, se aplicó el proceso para reducir el conjunto de datos a 2 variables, de las 8 que se tenia en un principio, esto se realizo debido a que se tenían variables con un alta correlación entre ellas, como en el caso del promedio con el porcentaje de asignaturas aprobadas en un mismo periodo; así también las calificaciones entre periodos guardan una alta correlación, con esto se busco reducir la complejidad y tiempo de construcción del modelo. Finalmente se realizó una normalización de valores de las dos variables generadas por el proceso anterior. Las 2 variables resultantes tienen un porcentaje de varianza explicada del 78 %, este porcentaje describe la cantidad de información que retiene el conjunto de datos para las dos dimensiones resultantes. Según Hair Junior y cols. (1998) el porcentaje mínimo aceptable de varianza explicada es 60 %.

3.4.2.1. Selección de datos para el análisis

Se consideraron todos los atributos de entrada para la construcción del modelo, estos se detallan en la Tabla 12. Luego de aplicado el proceso de reducción de dimensiones se tienen los siguientes atributos, descritos en la Tabla 13.

Conjunto de datos, experimento 2	
Campo	Descripción
1	Promedio en el primer ciclo de estudios
2	Promedio en el segundo ciclo de estudios
3	Promedio en el tercer ciclo de estudios
4	Promedio en el cuarto ciclo de estudios
1pro	Porcentaje de asignaturas reprobadas en el primer ciclo de estudios
2pro	Porcentaje de asignaturas reprobadas en el segundo ciclo de estudios
3pro	Porcentaje de asignaturas reprobadas en el tercer ciclo de estudios
4pro	Porcentaje de asignaturas reprobadas en el cuarto ciclo de estudios
5pro	Porcentaje de asignaturas reprobadas en el quinto ciclo de estudios

Tabla 12: Atributos seleccionados para el modelado

Conjunto de datos, experimento 2	
Campo	Descripción
avg1	Dimensión 1
avg2	Dimensión 2
5pro	Porcentaje de asignaturas reprobadas en el quinto ciclo de estudios

Tabla 13: Atributos obtenidos mediante PCA

3.4.3. Modelado

En la presente sección se describe que algoritmos se usaron, las pruebas de validación realizadas sobre los modelos y los resultados obtenidos. El código usado para la construcción del modelo se indica en el Anexo **??**

3.4.3.1. Técnicas usadas

Para la construcción del modelo, se realizó un proceso de clusterizacion basado en los 2 atributos descritos anteriormente (avg1, avg2), mediante dos algoritmos: KMeans y Affinity-Propagation; con los datos agrupados, se calculó un parámetro característico para cada cluster,

este parámetro es la relación entre el número total de registros en un cluster y el número de registros que reprobaron una asignatura en el quito periodo. Para realizar predicciones sobre nuevos registros, se tiene que pasar por el proceso se clusterizacion, en donde se le asigna un cluster basado en la semejanza de rendimiento académico, y por consiguiente se tiene el porcentaje de probabilidad de reprobar una asignatura, reflejado en el parámetro característico calculado.

3.4.3.2. Generación de pruebas modelo

Para la validación del modelo se uso las métricas de Brier Score loss, esta puntuación promedia el cuadrado de la diferencia entre el porcentaje de probabilidad y el resultado real (como uno o cero). Cuanto más baja sea la métrica, más precisa será la predicción **??**, y el área bajo la curva ROC (Receiver Operating Characteristic, o Característica Operativa del Receptor) La curva ROC es una medida de rendimiento para el problema de clasificación en diferentes configuraciones de umbrales. ROC es una curva de probabilidad y el area bajo la curva representa el grado o la medida de la separabilidad. Indica cuánto modelo es capaz de distinguir entre clases. Cuanto más alto es el AUC, mejor es el modelo en predecir 0s como 0s y 1s como 1s; se dividió el conjunto de datos en dos partes: 70 % para entrenamiento y 30 % para la validación del modelo.

3.4.3.3. Construcción del modelo

Como paso previo a la ejecución del proceso de agrupamiento (clusterizacion), es necesario determinar el número de clusters en los que se va a dividir el conjunto de datos, para esto se realizó el agrupamiento con dos clusters hasta veinte clusters, para cada uno estos agrupamientos se calculó la suma de las distancias medias de las muestras de cada cluster, esto se puede ver en el Gráfico 15 donde en el eje horizontal se encuentran los agrupamientos con di-

ferentes número de clusters y en el eje vertical la suma de las distancias entre muestras; como siguiente paso se debe identificar el punto donde la variación de la suma se mínima entre dos agrupaciones consecutivas. En este caso se determino que a partir de 8 clusters la agrupación no tiene variación significativa en la suma de la distancia de los clusters, y se ejecuto el proceso de clusterizacion con este parámetro, los resultados se pueden ver en el Gráfico 16.

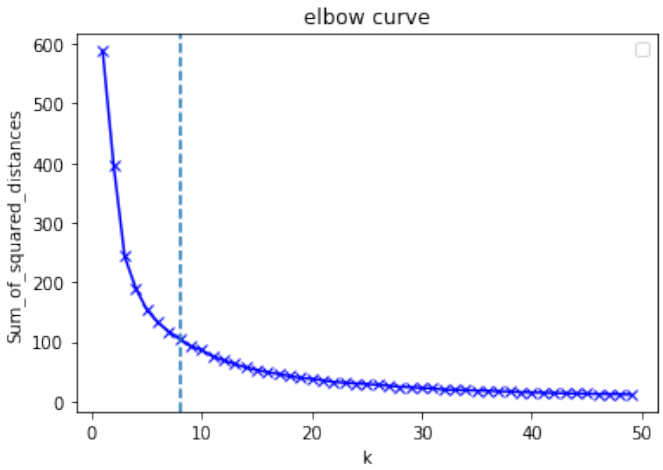


Figura 15: Elección del número de clusters para k-means

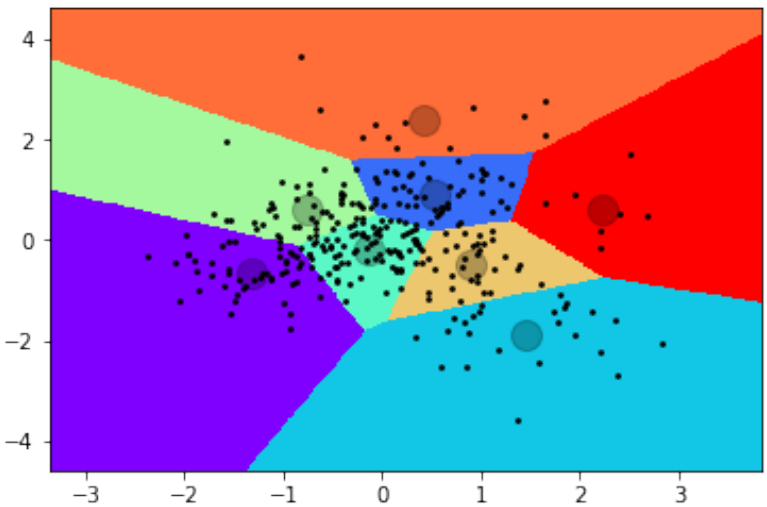


Figura 16: Visualización de datos agrupados con k-meas y 8 clusters.

Para el algoritmo de AffinityPropagation no es necesario determinar previamente el número de clusters deseados, el algoritmo internamente establece cual es el número de clusters óptimo, en este caso al algoritmo determino que la agrupación se de en 31 grupos. En el Gráfico 17 se puede observar como quedaron agrupados los datos. Con el fin de mantener consistencia entre los dos métodos usados para la clusterizacion, se ejecuto nuevamente el algoritmo K-means con 31 clusters; en resumen se tiene tres métodos para la agrupación de datos: k-means con 8 clusters, AffinityPropagation con 31 clusters y k-means con 31 clusters.

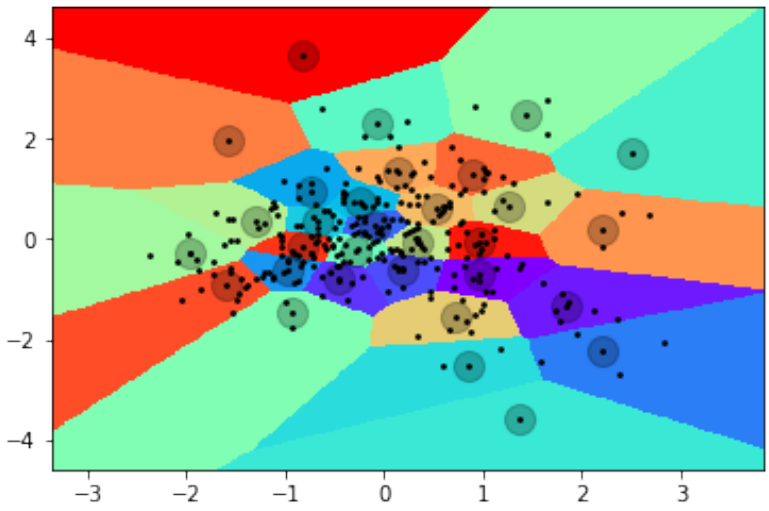
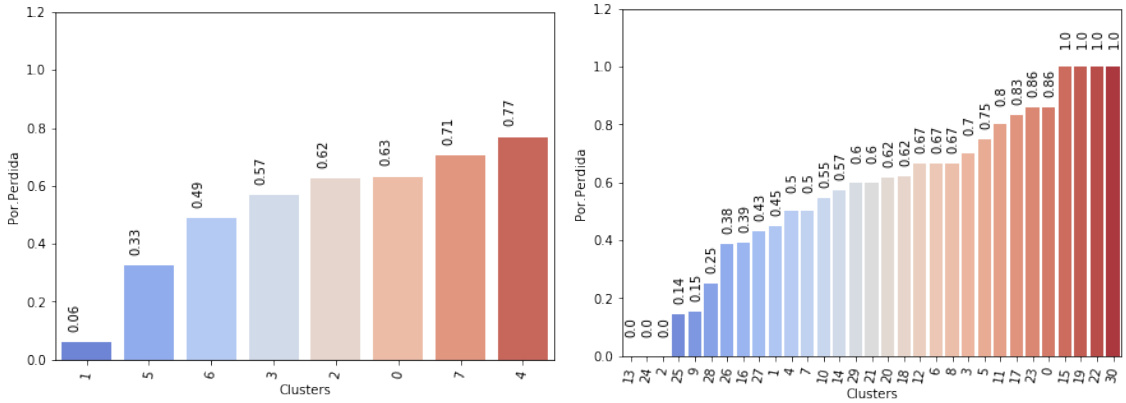


Figura 17: Visualización de datos agrupados con AffinityPropagation.

Luego de realizado el proceso de clusterizacion, se calculó el porcentaje de estudiantes que reprobaron una materia por cada cluster. En la Figura 19 se puede ver los resultados para el algoritmos k-means, en la Figura 18 para el algoritmo propagation affinity. En el eje horizontal se tiene a los diferentes clusters y en el vertical el porcentaje de registros que pierden una asignatura en el quinto ciclo.



(a) 8 clusters

(b) 31 clusters

Figura 19: Porcentaje de reprobados k-means

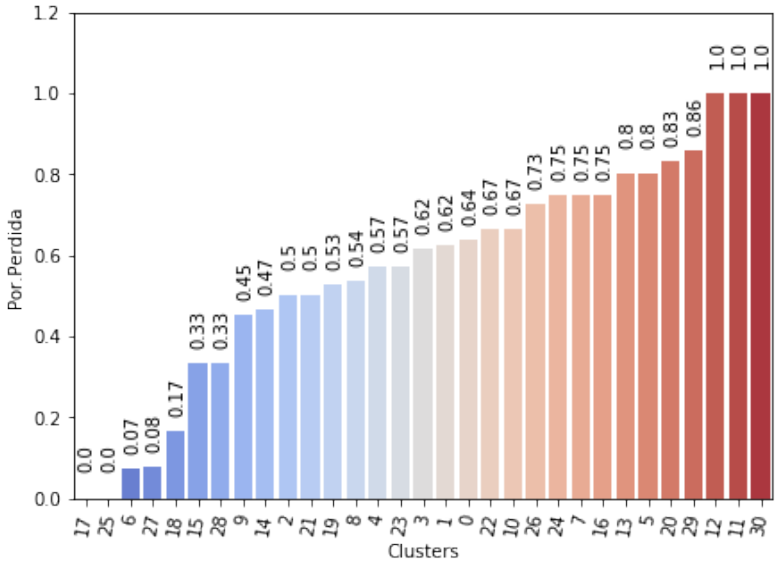


Figura 18: Porcentaje de reprobados Propagation affinity 31 clusters

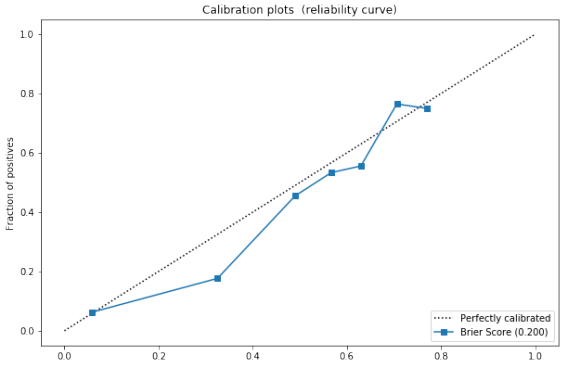
3.4.3.4. Evaluación del modelo

Para la evaluación del modelo se separó 126 registros, 67 pertenecientes a la clase 0 (no reprobó) y 59 a la clase 1 (reprobó), y se calculó la variable de respuesta usando el campo del

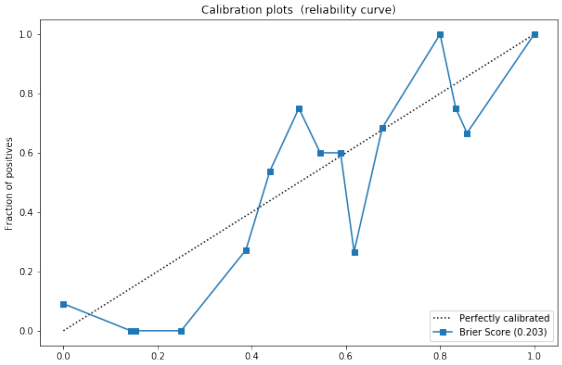
Algoritmos Agrupamiento		
Algoritmo	Brier Score	AUC
Propagation affinity	0.195	0.76
K-means 8 clusters	0.2	0.75
K-means 31 clusters	0.203	0.73

Tabla 14: Puntaje de los tres métodos de clusterizacion

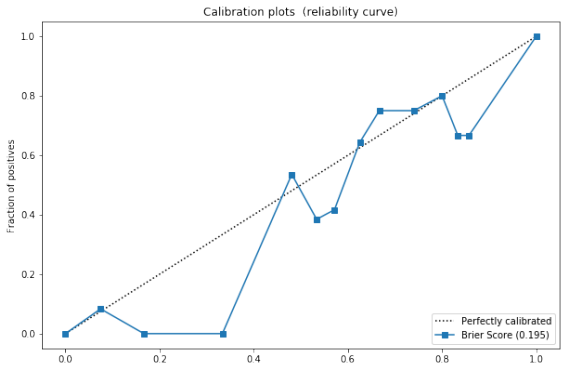
porcentaje de asignaturas aprobadas en 5 ciclo, si el estudiante reprobó una asignatura, la variable toma el valor de 1; si aprobó todas, toma el valor de 0, luego se procedió a predecir el cluster al que pertenecen y se comparó el valor del porcentaje de perdida contra la variable de respuesta. Los puntajes obtenidos por cada uno de algoritmos se describen en la Tabla 14, donde no se puede distinguir claramente que uno de los métodos sea el mejor, por esto se optó por realizar un gráfico de calibración en el cual en el eje horizontal están las probabilidades predichas por el modelo y en el eje vertical se especifica la frecuencia en la que el fenómeno ocurrió en el subconjunto de datos; un modelo de clasificación perfecto, describe una linea diagonal de 45 grados respecto al eje , esto quiere decir que si se cuentan todas las veces que un clasificador perfecto predijo que un fenómeno tiene 40 % de probabilidad de ocurrir, el porcentaje de acierto también debe ser del 40 %, es deseable que los presentes modelos sigan esta tendencia, es decir que los puntos calculados estén cerca de un clasificador perfecto. En este contexto se pudo observar que el algoritmo que mejor describe esta conducta es k-means de 8 clusters (Figura 20a), seguido de propagation affinity (Figura 20c) y finalmente k-meas de 31 clusters (Figura 20b).



(a) Calibration plot para k-means 8 clusters



(b) Calibration plot para k-means 31 clusters



(c) Calibration plot para Propagation affinity 31 clusters

3.5. Experimento 3 : Predicción de reprobación de asignaturas especificas, al momento de matriculación, basado en el datos académicos históricos

En esta sección se presenta el proceso de la construcción de un modelo predicativo, se parte con l preparación de los datos de entrada, luego se describe el proceso de transformación de los datos y finalmente la construcción y evaluación del modelo generado.

3.5.1. Preparación de los datos

Se presentan los datos a usar, el proceso de calculo de los atributos especiales, también se describe los métodos usados para la limpieza y filtrados de datos. El proceso ETL realizado para la transformacion de datos se indica en el Anexo **??**.

3.5.1.1. Datos para el análisis

En este experimento se utilizó los datos académicos de asignaturas cursadas desde el año 2008, ya sea aprobadas o reprobadas, de todas las carreras de la Universidad de Cuenca. Esto incluye número de créditos, número de matrícula, nota final, periodo en el que curso, asignaturas equivalentes y prerrequisitos de asignaturas.

3.5.1.2. Transformación de los datos

Como primer paso, se obtuvieron las matrículas de los estudiantes que tienen asignaturas con estado de aprobación aprobado y reprobado, aquí se tenia 2569575 registros, luego de esto se cruzaron los datos con mallas curriculares y asignaturas por nivel para saber el número de

créditos de cada asignatura, aquí se redujeron los registros a 1914684 debido a que no todas las matrículas contaban con la información de la estructura curricular a la que pertenecen, finalmente se eliminaron algunos registros que no contaban con una identificación de persona valido, quedando 1914669 registros.

En base a los datos obtenidos se procedió a calcular el promedio acumulado por cada periodo de estudios de los estudiantes, es decir, calcular el promedio de calificaciones que tuvieron hasta el periodo anterior que tomaron una asignatura. Esto se realizó usando la librería pandas existente dentro de python.

Otro atributo que se calculó fue el de la dificultad que tiene una asignatura, esto es el promedio ponderado por el número de créditos de las calificaciones de todos los estudiantes que han tomado la asignatura o su equivalente. La fórmula de cálculo esta dada por:

$$Difficulty_c = \frac{\sum_{t \epsilon BE_c} \sum_{j=1}^{m_t} G_{j,t} * W_t}{\sum_{t \epsilon BE_c} W_t * m_t} \tag{3.1}$$

donde c, curso actual; t, curso equivalente al actual; BE_c, el conjunto de cursos equivalentes al actual c; m_t, número total de estudiantes en el curso t; $G_{j,t}$, la calificación del jvo estudiante en el curso t ;W_t, el número de créditos del curso t (Vialardi y cols., 2011).

Para considerar nuevas calificaciónes de estudiantes, este atributo deber ser recalculado antes del inicio de cada periodo. Como el valor de este atributo es proporcional al promedio de calificaciones de los estudiantes que se han matriculado en el curso, un valor menor representa un curso mas difícil (Vialardi y cols., 2011). Este atributo fue calculado mediante un proceso ETL (extracción, transformación y carga), en la herramienta de Pentaho, se tomó como entrada los datos obtenidos en la primera parte del experimento, luego se recuperaron las asignaturas equivalentes para realizar un agrupación de asignaturas iguales y realizar los cálculos requeridos por la formula utilizada al final se obtuvieron 6262 registros que contienen la identificación

de la asignatura y su dificultad.

El siguiente atributo calculado fue el de potencial, el cual representa que tan competente es un estudiante para cursar una asignatura basado en las calificaciones obtenidas en las asignaturas que son prerrequisitos de la actual, este atributo es calculado como un promedio ponderado por el número de créditos de la calificaciones en asignaturas prerrequisitos dividido para sus respectivas dificultades. El potencial esta representado por:

$$Potential_{s,c,d} = \frac{\sum_{t \epsilon SPC_{c,d}} \sum_{v=1}^{H_t} (\frac{G_{s,t,v} * W_t}{D_t})}{\sum_{t \epsilon SPC_{c,d}} W_t * H_t} \tag{3.2}$$

donde s, estudiante; c, curso actual; d, distancia para el calculó del potencial; t, curso prerrequisito; $SPC_{c,d}$, conjunto de prerrequisitos del curso c a una distancia d; H_t, número de veces que el estudiante se matriculo en un curso t; $G_{s,t,v}$, calificación del estudiante s en el curso t en el intento v; W_t, número de créditos del curso t; D_t dificultad del curso t(Vialardi y cols., 2011).

En el presente trabajo se tomó en cuenta solo los prerrequisitos directos de las asignaturas, entonces el valor para el parámetro d es 1, de acuerdo a la expresión, mientras más alto el valor del potencial, más alta sera la probabilidad de que el estudiante tenga un buen rendimiento en el curso(Vialardi y cols., 2011). En el caso de que la asignatura no tenga prerrequisitos, se calculara el potencial basado en todas las asignaturas cursadas por el estudiante hasta ese momento.

Al igual que el atributo anterior de dificultad, este fue calculado usando la herramienta de Pentaho, se implementó un proceso para el calculó de asignaturas que tienen prerrequisitos y otro para los que era necesario calcular el potencial basado en todas las asignaturas cursadas hasta ese momento. Fue necesario obtener de la base de datos del sistema de gestión académica

la información de prerrequisitos de asignaturas. En este proceso también se integró el parámetro de dificultad obtenido en el paso anterior, al final se obtuvieron los siguientes atributos: identificación de persona, identificación de asignatura, valor del potencial.

Una vez que se calcularon todos los parámetros requeridos, se procedió a integrarlos en 1 solo conjunto de datos y así proceder a la construcción del modelo, en este punto de igual manera se calculó la variable de respuesta, que contiene dos clases: aprobó (1) o reprobó (0), al finalizar la transformación e integración de datos se cuenta con los atributos descritos en la Tabla 15 para pasar a la siguiente etapa.

Conjunto de datos, experimento 3	
Campo	Descripción
id_estructura_curricular_x_estudiante	identificador de estudiante por estructura
id_servicio_asignatura	identificador de asignatura
nombre	nombre de asignatura
número	nivel de asignatura
id_periodo	periodo que cursa
créditos	número de créditos de asignatura
veces_cursa	número de matrículas
dificultad	parámetro calculado
potencial	parámetro calculado
promedio_periodo	promedio acumulado hasta el periodo anterior
créditos_por_periodo	créditos cursados el periodo actual
clase_respuesta	aprueba o reprueba

Tabla 15: Conjunto de datos luego de transformación e integración

3.5.1.3. Limpieza y filtrado de los datos

Una vez que se tuvieron todos los datos integrados en un solo conjunto de datos, se paso a realiza un proceso de eliminación de registros con valores nulos, estos fueron 4114 filas que no se tomaron en cuenta, puesto que, principalmente tenían valores nulos en el campo de la variable

de respuesta. Otro proceso que se llevo a cabo fue un balanceo de clases, en el cual se asigna un peso a cada registro en base a la clase de su variable respuesta para que cada una tenga el mismo peso a la hora de realizar la clasificación, esto fue necesario debido a que el conjunto de datos tenia 850 % de registros para la clase 1 (aprueba) y 15 % para la clase 0 (reprueba). Otro proceso que se considero necesario fue la normalización de los valores de todos los atributos numéricos, para reducir la dispersión de valores y con esto disminuir la complejidad del modelo resultante.

3.5.1.4. Selección de datos para el análisis

Luego de tener los datos preparados para la construcción del modelo se eliminaron los atributos que corresponden a identificadores de la base de datos y con esto el conjunto de datos quedó con los siguientes campos descritos en la Tabla 16 y en total se contó con 207158 registros para la clase 0 y 1559547 para la clase 1. Los atributos seleccionados para este experimento se basan en el trabajo de Vialardi y cols. (2011).

Conjunto de datos finales, experimento 3	
Campo	Descripción
nombre	nombre de asignatura
número	nivel de asignatura
créditos	número de créditos de asignatura
veces_cursa	número de matrículas
dificultad	parámetro calculado
potencial	parámetro calculado
promedio_periodo	promedio acumulado hasta el periodo anterior
créditos_por_periodo	créditos cursados el periodo actual
clase_respuesta	aprueba o reprueba

Tabla 16: Conjunto de datos luego de limpieza de datos

3.5.2. Modelado

En esta sección se describe los algoritmos usados para la construcción del modelo y se define la forma de evaluación.

3.5.2.1. Técnicas a usar

Los dos algoritmos a usar son: J48, debido a que es uno de los algoritmos mas usados en el campo de la minería de datos académica (Vialardi y cols., 2011) y el algoritmo Naive Bayes que se lo usa generalmente como base para la comparación de rendimiento de la clasificación.

3.5.2.2. Generación de pruebas modelo

Para la validación del modelo se reservo el 30 % de registros de cada clase, mediante la extracción de una muestra aleatoria, como métricas para escoger el mejor algoritmo se uso el porcentaje de aciertos.

3.5.2.3. Construcción del modelo

Para la construcción del modelo del presente experimento, se usó la herramienta WEKA debido a que tiene soporte para atributos categóricos o nominales dentro de sus algoritmos de clasificación, en este caso necesario por la inclusión del atributo nombre de asignatura. Se realizó un proceso de optimización de parámetros para el algoritmo J48, para esto se lo ejecuto con diferentes configuraciones variando el valor del parámetro, número mínimo de muestras por hoja (10,20,40) ;de estos valores el que produjo un mejor resultado fue el valor 40 muestras, se alcanzó el resultado descrito en la Figura 21 usando la prueba de t pareada con un nivel de significancia de $0{,}05$. Ahí se puede ver listadas las tres configuraciones usadas: la primera (1) con valor 10, la segunda (2) con valor 20, la tercera (3) con un valor de 40; si se toma como

base la configuración (3) se puede ver que existe una diferencia significativa con las demás configuraciones, es decir, que la configuración (3) es la que produce el mejor resultado en cuanto al porcentaje de aciertos, esto esta denotado por el símbolo $*$ que indica los resultados fueron peores que la base.

```
Dataset                 (3) trees.J4 | (1) trees (2) trees
----------------------------------------------------------
'todo_3-weka.filters.unsu  (1)   81.82 |   80.80 *   81.46 *
----------------------------------------------------------
                              (v/ /*) |   (0/0/1)   (0/0/1)


Key:
(1) trees.J48 '-C 0.25 -M 10' -217733168393644444
(2) trees.J48 '-C 0.25 -M 20' -217733168393644444
(3) trees.J48 '-C 0.25 -M 40' -217733168393644444
```

Figura 21: Prueba t pareada para el número minino de muestras por hoja

En base a los resultados anteriores, se toma como valor óptimo para el número mínimo de muestras en las hojas a 40, con esto ahora se procede a optimizar el parámetro de factor de confianza, mientras mas bajo sea este valor mayor sera la poda(pruning) que se realice al árbol, los valores usados fueron (0.01,0.05,0.1,0.25). El resultado del test t pareado se puede observar en la Figura 22, ahí se pude notar que el parámetro óptimo es para la configuración (2), correspondiente al valor de parámetro 0.05, con un porcentaje de aciertos de 81.99 %, las configuraciones restantes tuvieron una diferencia significativa y produjeron un peor modelo de clasificación.

```
Dataset                 (2) trees.J4 | (1) trees (3) trees (4) trees
----------------------------------------------------------------------
'todo_3-weka.filters.unsu  (1)   81.99 |   80.64 *   81.97 *   81.82 *
----------------------------------------------------------------------
                              (v/ /*) |   (0/0/1)   (0/0/1)   (0/0/1)


Key:
(1) trees.J48 '-C 0.01 -M 40' -217733168393644444
(2) trees.J48 '-C 0.05 -M 40' -217733168393644444
(3) trees.J48 '-C 0.1 -M 40' -217733168393644444
(4) trees.J48 '-C 0.25 -M 40' -217733168393644444
```

Figura 22: Prueba t pareada para el factor de confianza

Finalmente se comparó los algoritmos de Naive Bayes y J48 con parámetros óptimos, los resultados se pueden ver en la Figura 23, aquí se puede notar que el algoritmo j48 tiene una diferencia significativa con Naive Bayes, con respecto al porcentaje de aciertos.

```
Dataset                   (1) trees.J4 | (2) bayes
----------------------------------------------------
'todo_3-weka.filters.unsu  (1)   81.99 |   76.40 *
----------------------------------------------------
                               (v/ /*) |   (0/0/1)


Key:
(1) trees.J48 '-C 0.05 -M 40' -217733168393644444
(2) bayes.NaiveBayes '' 5995231201785697655
```

Figura 23: Comparación entre J48 y Naive Bayes

3.5.2.4. Evaluación del modelo

En la Figura 24, se puede ver el reporte de la clasificación realizadas con el algoritmo j48 y sus parámetros optimizados, aquí se destaca el porcentaje de acierto que esta ubicado en 83.142 %, este modelo fue validado mediante un subconjunto de datos extraído previo a la clasificación que contiene 530012 registros, este subconjunto de datos de validación mantiene

la misma proporción de registros entre las clases de la variable respuesta que el subconjunto de datos de clasificación.

```
=== Summary ===

Correctly Classified Instances      440663               83.1421 %
Incorrectly Classified Instances     89349               16.8579 %
Kappa statistic                          0.4454
Mean absolute error                      0.241
Root mean squared error                  0.3463
Relative absolute error                 48.1922 %
Root relative squared error             69.2536 %
Total Number of Instances           530012

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.821    0.167    0.395      0.821    0.533      0.490    0.893     0.637     0
                 0.833    0.179    0.972      0.833    0.897      0.490    0.893     0.979     1
Weighted Avg.    0.831    0.177    0.905      0.831    0.854      0.490    0.893     0.938

=== Confusion Matrix ===

      a      b   <-- classified as
  51039  11108 |      a = 0
  78241 389624 |      b = 1
```

Figura 24: Reporte de clasificación algoritmo j48

3.6. Despliegue

Se describe el proceso a seguir para poner los resultados de los modelos a disposición de los actores interesados.

3.6.1. Arquitectura del sistema de recomendación

En la Figura 25, se presenta el flujo de datos propuesto para la implementación del sistema de recomendación para estudiantes y autoridades, este será el encargado de presentar los datos resultantes de los modelos generados en los experimentos anteriores.

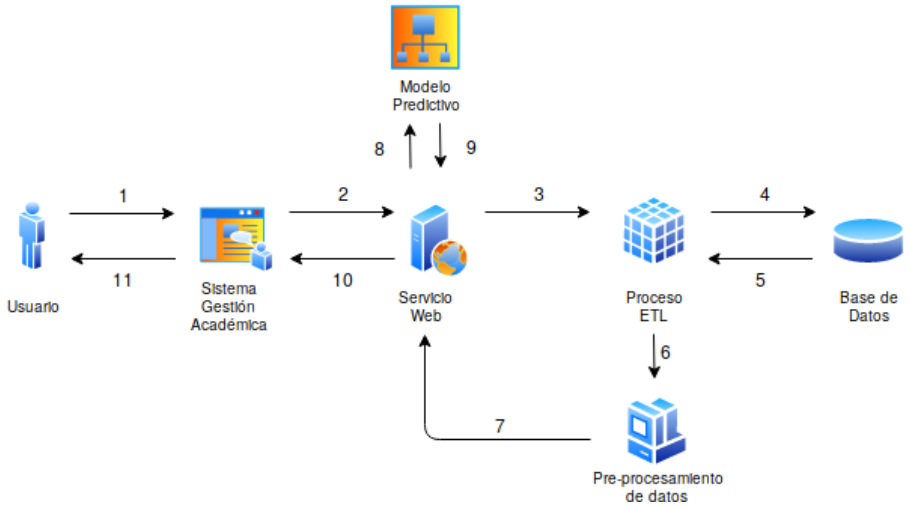


Figura 25: Flujo de datos para presentación de resultados

A continuación se describe el proceso a seguir para la presentación de datos. El estudiante ingresa en el sistema de gestión académica (SGA), para realizar su matrícula al inicio de un periodo (1), aquí el SGA llama al servicio web, enviando los siguientes datos del estudiante:identificación, periodo y asignatura (2); el servicio web ejecuta un proceso ETL para extraer la información necesaria para la ejecución de los modelos(3), el proceso ETL ejecuta consultas SQL para obtener la información del estudiante (4) y la devuelve al ETL (5), que a su vez envía los datos a un script en python para realizar una limpieza y filtrado de datos (6) y envían el resultado al servicio web (7); el servicio invoca a los modelos predictivos (8) se devuelven los datos inferidos y probabilidades (9), el servicio web envía estos datos al SGA para su presentación en la interfaz (10) y finalmente los datos son presentados al usuario (11).

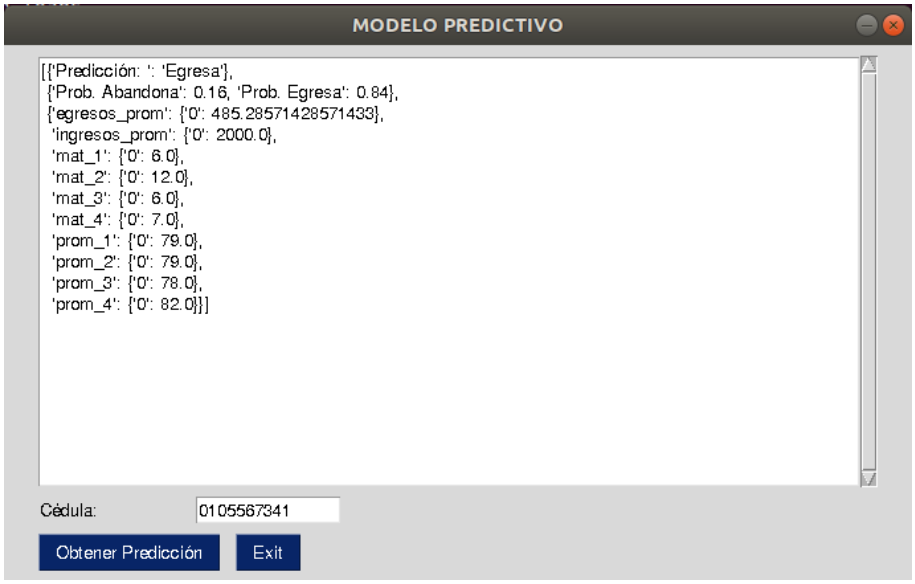


Figura 26: Interfaz aplicación prototipo

En el presente trabajo se realizó una implementación inicial; realizando el proceso ETL, la consulta a base de datos y el pre-procesamiento de datos para todos los estudiantes que están cursando al menos el cuarto ciclo de estudios; estos datos ya preprocesados se guardaron en una base de datos, se implementó una aplicación de demostración que, en base a estos datos ya procesados realiza consultas a los modelos predictivos guardados en disco, y envía la respuesta sobre la clase predicha y su probabilidad. En la Figura 26 se puede ver la interfaz construida para la aplicación prototipo. En el Anexo **??** se indica el código usado para la creación de la aplicación y su interfaz.

Capítulo IV

DISCUSIÓN DE RESULTADOS

Teniendo en cuenta los objetivos de la minería de datos planteados en la Sección 3.1.3, se puede decir que se han cumplido con todos. Cada uno de los modelos obtenidos en los tres experimentos realizados han tenido un porcentaje de acierto sobre el 70 %, luego de haber determinado las mejores configuraciones para el conjunto de datos. Para la evaluación de los modelos se empleo el método de validación cruzada, que permite verificar la independencias entre los subconjuntos de datos de entrenamiento y validación, de manera que se puede estar seguro que el modelo tendrá un buen desempeño para nuevos registros. Siguiendo con esta linea, se emplearon curvas de aprendizaje para comprobar que el modelo no sufra de sobre-ajuste, ajustando los parámetros donde fue necesario para evitar este fenómeno y obtener un modelo con mejor generalización.

En cuanto al modelo resultante del experimento 1, se puede ver que tanto para la clase deserta como para la clase aprueba se tiene bueno valores de precisión, esto quiere decir que el modelo generado puede predecir de manera satisfactoria ambas clases. Esto se logró mediante la aplicación de técnicas de sub-muestreo para balancear las clases antes de su procesamiento

por el algoritmo de clasificación.

Para el experimento 2, se uso un gráfico de calibración para obtener el mejor método de clusterizacion, ahí se pudo notar un mejor desempeño de la predicción de probabilidad cuando se dividieron los datos en menos grupos(clusters), esto puede ser debido a que mientras mas clusters hay, mas especifico se vuelve el modelo y pierde su capacidad de predecir nuevos valores. Cabe destacar que antes de proceder con la clusterizacion se ejecuto un proceso de reducción de dimensionalidad del conjunto de datos, con el objetivo de reducir la complejidad del modelo y agrupar atributos con una alta correlación. Con esto se demostró la factibilidad de realizar este tipo de análisis para las demás carreras y poder extenderlo para todos los periodos de estudio.

Finalmente con el experimento 3, la mayor parte de trabajo fue puesto en la recopilación, limpieza y transformación de los datos, se usaron dos atributos calculados en base al desempeño histórico de una asignatura y el desempeño en asignaturas prerrequisitos, además se calculó el promedio acumulado de cada estudiante hasta el periodo anterior de estudios, se probo con dos algoritmos de clasificación: j48 que crea un árbol de decisión del cual se pueden obtener patrones y así saber en que condiciones los estudiantes aprueban o reprueban una determinada asignatura; y Naive Byaes que se lo uso como base para la comparación del algoritmos. Se ejecutaron procesos de balanceo de clases, normalización de datos y optimización de parámetros para alcanzar un porcentaje de aciertos del 83 %, la validación se la realizó extrayendo un conjunto de datos de prueba con la misma proporción de registros para cada clase que el conjunto de entrenamiento.

Capítulo V

CONCLUSIONES

5.1. Conclusiones

Los modelos obtenidos en el presente trabajo, presentan la suficiente precisión como para ser usados en un escenario real, en el cual será necesario medir su efectividad antes de uso general, antes de llegar a los modelos fue necesario realizar un proceso exhaustivo de análisis, selección,transformación y filtrado de datos, esta fue la parte mas ardua del trabajo debido a que se accedió directamente a las bases de datos transaccionales, y no se uso un datawarehouse debido a que al inicio del trabajo estaba en una etapa temprana de desarrollo. Se probó además varios algoritmos de minería de datos con el objetivo de encontrar el que mejor se adapta a los datos. Dentro de la selección de atributos que influyen en la deserción académica se concluye, luego del análisis realizado, que los campos relacionados al rendimiento académico son los más relevantes para los modelos de predicción, seguido por los ingresos y egresos económicos; los demás campos relacionados al factor socio-económico no se mostraron como un factor decisivo.

Se verifica el cumplimiento de los objetivos planteados al inicio del trabajo:

- Realizar un proceso de análisis de variables y fuentes de datos que intervienen en el rendimiento académico. En el experimento 1 se realizó un análisis de variables para encontrar cuales son las que tienen mas influencia en la deserción académica, además para los experimentos 1,2,3 se analizaron las bases de datos y se identificaron que tablas y campos contienen variables para realizar predicción de reprobación de asignaturas.

- Obtener varios modelos, conformados por reglas o patrones que permitan la clasificación o predicción de estudiantes en diferentes categorías de rendimiento académico o estado académico (cursando o deserto). En cada experimento se obtuvo al menos 2 modelos diferentes, en los cuales se identificó mediante métricas cuales fueron los mejores para cada objetivo. Se logró ir desde lo general, que es, si un estudiante termina su carrera, pasando por si reprueba en un periodo, hasta tener específicamente para cada asignatura si aprueba o no.

- Validar la precisión de los modelos generados. Se usaron métricas para validación del rendimiento de los modelos, se especificaron los escenarios de pruebas, con validación cruzada para tener una mejor generalización para nuevos datos, todos los modelos obtenidos tuvieron un porcentaje mayor al 70 %.

  Proponer un sistema prototipo que permita la visualización de datos y predicciones. Se especifico la secuencia de datos que se puede seguir al momento de consultar los datos en el Sistema de Gestión Académica.

En cuanto a las preguntas de investigación:

Pregunta 1: ¿Se puede predecir si un estudiante va a egresar o abandonar sus estudios, con una precisión razonable, en una etapa temprana de su carrera; basado en sus calificación

y/o situación socio-económica? Si se puede predecir el abandono o egreso de un estudiante con precisión del 73 % al momento que finaliza el cuarto ciclo de estudios.

Pregunta 2: ¿Se puede predecir si un estudiante va a reprobar al menos una asignatura, con una precisión razonable, al inicio de cada periodo académico; basado en sus calificación anteriores? Si se puede predecir la reprobación en al menos una asignatura, con una precisión del 75 % al momento de finalizar un periodo

Pregunta 2: ¿Se puede predecir si un estudiante va a reprobar una asignatura especifica, con una precisión razonable, al inicio de cada periodo académico; basado en sus calificación anteriores? Si se puede predecir, con una precisión del 83 %, al momento de la matrículación.

5.2. Futuras líneas de investigación

Para futuros trabajos es recomendable la implementación de los modelos para la validación por parte de los estudiantes y autoridades en escenarios reales. Se recomienda la implementación del experimento 2 para todas las carreras de la universidad y todos los periodos de estudios, en este trabajo no se realizó, debido a que se requiere realizar un análisis personalizado para cada carrera y cada periodo, y así obtener buenos resultados. Además se aconseja la inclusión de datos de nuevos sistemas que van teniendo un mayor uso en la institución, como, gestión de sílabos, evaluación al docente y plataforma virtual de aprendizaje.

REFERENCIAS

Artunduaga Murillo, M. (2008). Variables que influyen en el rendimiento académico en la universidad. *Madrid, España*.

CACES. (2018a). Institucional. *CACES*. Descargado July 2, 2018, de https://www.caces.gob.ec/web/ceaaces/institucional

CACES. (2018b). Modelo de evaluación institucional de universidades y escuelas politécnicas 2018 version preliminar. , 24-25.

Eckert, K. B., y Suénaga, R. (2015). Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos. *Formación universitaria*, *8*(5), 03–12.

Hair Junior, J. F., Anderson, R. E., Tatham, R. L., y Black, W. C. (1998). Multivariate data analysis. *New Jersey*.

Han, J. (2011). *Data mining : concepts and techniques*. Burlington: Elsevier Science.

Mariscal, G., Marban, O., y Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166.

Meinshausen, N., y Bühlmann, P. (2010, julio). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473. Descargado de https://doi.org/10.1111/j.1467-9868.2010.00740.x doi: 10.1111/j.1467-9868.2010.00740.x

Mohamad, S. K., y Tasir, Z. (2013, noviembre). Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, *97*, 320–324. Descargado de https://doi.org/10.1016/j.sbspro.2013.10.240 doi: 10.1016/j.sbspro.2013.10.240

Ordoñez Briceño, K. F. (2013). *Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la modalidad abierta y a distancia de la utpl.* (Tesis Doctoral no publicada). UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA.

Pereira, R. T. (2010). Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos. *Revista Guillermo de Ockham*, *8*(1).

Pereira, R. T., Romero, A. C., y Toledo, J. J. (2013). Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. *Revista vínculos*, *10*(1), 373–383.

Riquelme Santos, J. C., Ruiz, R., y Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.*.

Romero, C., y Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

Ruiz, C. G., Marcela, D. M. D., Gallego, J. F., Castano, V. E., Gomez, S. G., y Portilla, K. G. (2009). *Desercion estudiantil en la educacion superior colombiana: metodologia de seguimiento, diagnostico y elementos para su prevencion*. Ministerio de Educacion Nacional.

Sa, C. L., Hossain, E. D., bin Hossin, M., y cols. (2014). Student performance analysis system (spas). En *The 5th international conference on information and communication technology for the muslim world (ict4m)* (pp. 1–6).

Sieminski, A., Kozierkiewicz, A., Nunez, M., y Ha, Q. T. (Eds.). (2018). *Modern approaches for intelligent information and database systems*. Springer International Publishing. Descargado de https://doi.org/10.1007/978-3-319-76081-0 doi: 10.1007/978-3-319-76081-0

Spositto, O., Etcheverry, M., Ryckeboer, H., y Bossero, J. (2010). Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. En *Novena conferencia iberoamericana en sistemas, cibernética e informática, cisci* (Vol. 29, pp. 06–2).

Timarán, R., Calderón, A., y Jiménez, J. (2013). Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. *Ventana informática*, *28*, 31–47.

Tomek, I. (1976). Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, *6*, 769–772.

Torres, C. Z., Ramos, C. A., y Moraga, J. L. (2016). Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos. *Ciencia Amazónica:(Iquitos)*, *6*(1), 73–84.

Tukey, J. W. (1977). *Exploratory data analysis*. Pearson. Descargado de https://www.xarg.org/ref/a/0201076160/

Universidad de Cuenca. (2017). Plan estratÉgico de desarrollo institucional 2017 - 2021.

Vera, C. M., Morales, C. R., y Soto, S. V. (2012). Predicción del fracaso escolar mediante técnicas de minería de datos. *Revista Iberoamericana de Tecnologías del/da Aprendizaje/Aprendizagem*, *109*.

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., y Ortigosa, Á. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User modeling and user-adapted interaction*, *21*(1-2), 217–248.

Witten, I. H. (2005). *Data mining : practical machine learning tools and techniques*. Amsterdam Boston, MA: Morgan Kaufman.

Ye, N. (2003). *The handbook of data mining*. CRC Press.

Zhu, B., Deng, Q., y He, X. (2014). A new hybrid model of feature selection for imbalanced data. En J. Xu, J. A. Fry, B. Lev, y A. Hajiyev (Eds.), *Proceedings of the seventh international conference on management science and engineering management* (pp. 549–558). Berlin, Heidelberg: Springer Berlin Heidelberg.